

Pontificia Universidad Católica del Ecuador

Facultad De Ingeniería

Escuela de Sistemas



TEMA:

Desarrollo de un modelo de machine learning para predecir el rendimiento académico de los estudiantes universitarios en términos de aprobación o reprobación de materias utilizando la metodología CRISP-DM

AUTOR:

Sebastián Felipe Tamayo Proaño

TRABAJO PREVIA A LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN
TECNOLOGÍAS DE LA INFORMACIÓN

QUITO, 2023 – 2024

DEDICATORIA

En primer lugar, deseo dedicar este trabajo de titulación a cada distinguido lector que sienta curiosidad por los intrínsecos contenidos de este escrito, puedo asegurar que cada parte de este trabajo ha sido realizado poniendo cada parte de mí, mi amor y mis principios; todo este trabajo ha sido llevado a cabo con el inherente afán de que ustedes, estimados lectores, se deleiten y enriquezcan sus conocimientos con cada componente que podrán encontrar en las siguientes páginas de este extenso trabajo de titulación.

Adicionalmente, deseo dedicar este trabajo a todos aquellos que me han apoyado a lo largo de mi vida, en especial a mi familia, a mis padres, a mi hermana, a mis abuelos, a todos; estoy más que seguro que de no haber contado con ustedes, no fuera la persona que soy ahora. Por otra parte, agradezco a todas las personas con las que he tenido la oportunidad de compartir momentos, sean momentos efímeros, maravillosos o trágicos, todos aquellos que han formado parte de mi vida me han ayudado a forjarme y motivarme a tratar de ser la mejor versión de mí mismo, e intentar demostrarlo con este trabajo.

Finalmente, deseo dedicar este trabajo de titulación a mi persona, esto, como un regalo, como un hito que me permita recordar que a pesar de cuantos abismos existan en los diferentes caminos que progresivamente se han ido despejando a medida que he ido culminando este trabajo, siempre podré levantarme. Sebastián, si alguna vez vuelves a leer esto, no te rindas, no dejes morir a tus principios ni a tu ser.

AGRADECIMIENTO

En primer lugar, deseo agradecer a Dios, a los seres más allá del entendimiento, agradecerles por su guía y sabiduría, que, a pesar de que he sido necio y muy humano en la mayoría de las decisiones que he tomado en mi vida, su providencia logra avivar mi espíritu y fortaleza para cumplir mi misión en el mundo.

Adicionalmente, deseo agradecer a la Pontificia Universidad Católica del Ecuador por permitirme forjarme no solo como profesional sino como humano, en especial, quiero agradecer al Mgtr. Jorge Alejandro Alarcón Mena por su excelente labor como profesor y profesional, es un ejemplo por seguir y un excelente símbolo que mantiene el prestigio de la universidad ante el constante crecimiento de una eventual sombra de la decadencia.

Aprovecharé también para agradecer profundamente a mis amigos, especialmente a Sofía Villacís y Bryan Cuvi, con ellos he compartido toda mi vida universitaria y estoy profundamente agradecido por todo el apoyo que me han dado y la sincera amistad que hemos tenido.

Finalmente, deseo agradecer a mi tutor de trabajo de titulación, el Mgtr. Charles Escobar y al Ing. Xavier Vásquez, por su guía, paciencia y voluntad de apoyarme a concretar un trabajo de titulación que cumpla con las expectativas de sus futuros lectores.

RESUMEN

El presente proyecto de titulación aborda el desarrollo de un modelo de machine learning para la predicción del rendimiento académico de estudiantes universitarios enfocado en la Pontificia Universidad Católica del Ecuador. Este modelo tiene como objetivo práctico el predecir si un estudiante reprobará o aprobará alguna determinada asignatura dada el área de la misma, sin embargo, su objetivo holístico es el proporcionar una solución que permita ayudar a aquellos estudiantes que tienen alta de probabilidad de rendir negativamente en un periodo académico dado, lo que permitiría tanto al estudiante como a la universidad, emplear planes de acción para prevenir dicho resultado y disminuir la probabilidad de que este determinado estudiante posea un rendimiento desfavorable.

El trabajo de titulación fue realizado haciendo uso de la metodología CRISP-DM, la cual consta de 6 fases que incluye la comprensión del negocio, la comprensión de los datos, su preparación, modelado, evaluación y despliegue del modelo. De manera general, se ha realizado un análisis exhaustivo de los datos para determinar las mejores variables posibles para el modelo de predicción de rendimiento académico, posteriormente, se prepararon los datos, se desarrollaron varios modelos de machine learning y tras una evaluación de estos.

ÍNDICE

ÍNDICE DE FIGURAS, GRÁFICOS Y TABLAS	VII
ÍNDICE DE FIGURAS	VII
ÍNDICE DE TABLAS	LX
CAPÍTULO I: INTRODUCCIÓN	1
1. Introducción	1
1.1. Justificación	1
1.1.1. Alcance	3
1.2. Planteamiento del problema.....	3
1.3. Objetivos	4
1.3.1. Objetivo General.....	4
1.3.2. Objetivos Específicos.....	5
1.4. Estructura del trabajo	5
1.4.1. Definición de objetivos de la investigación y entendimiento del problema ...	6
1.4.2. Determinación del criterio de éxito y el resultado deseado	8
1.4.3. Identificación de las partes interesadas.....	9
1.4.4. Identificación de posibles riesgos y desafíos	10
1.5. Notas respecto a los códigos realizados en el trabajo	11
CAPÍTULO II: REVISIÓN SISTEMÁTICA DE LITERATURA	12

2.	Revisión sistemática de literatura	12
2.1.	Rendimiento académico en estudiantes universitarios	12
2.2.	Modelos de Machine Learning para la predicción del rendimiento académico	14
2.3.	Metodología CRISP-DM	15
2.3.1.	Entendimiento del negocio	16
2.3.2.	Entendimiento de los datos	17
2.3.3.	Preparación de los datos.....	18
2.3.4.	Modelado	18
2.3.5.	Evaluación.....	19
2.3.6.	Despliegue.....	19
	CAPÍTULO III: MARCO TEÓRICO.....	20
3.	Marco Teórico.....	20
3.1.	Aprendizaje supervisado y no supervisado.....	20
3.2.	Algoritmos de machine learning utilizados en el estudio	20
3.2.1.	Regresión Logística	21
3.2.2.	Árbol de decisión	21
3.2.3.	Random Forest	22
3.2.4.	Support Vector Machines (SVM)	22
3.2.5.	Redes Neuronales.....	22
3.3.	Métricas de evaluación de modelos de machine learning.....	23

3.3.1. Precisión (Accuracy).....	23
3.3.2. Recall o Sensibilidad.....	23
3.3.3. Precisión (Precision)	24
3.3.4. F1-Score.....	24
3.3.5. Curva ROC y AUC	24
CAPÍTULO IV: METODOLOGÍA.....	25
4. Metodología	25
4.1.Comprensión de los datos	25
4.1.1. Librerías requeridas para la fase	26
4.1.2. Entendimiento de los conjuntos de datos obtenidos	27
4.1.3. Carga de los conjuntos de datos.....	29
4.1.4. Verificación de consistencia entre periodos de los conjuntos de datos	36
4.1.5. Identificación de las variables de los conjuntos de datos	43
4.1.6. Identificación de la información personal identificable.....	52
4.1.7. Anonimización de datos.....	58
4.1.8. Identificación de tipos de variables.....	66
4.1.9. Selección preliminar de variables relevantes	121
4.1.10. Exploración de variables numéricas	135
4.1.11. Exploración de variables categóricas.....	203
4.1.12. Exploración de relaciones	280

4.1.13.	Selección avanzada de variables relevantes.....	390
4.1.14.	Vista general de los conjuntos de datos finales	397
4.1.15.	Exportación de los conjuntos de datos para la siguiente fase	402
4.1.16.	Planteamiento de la hipótesis para el modelo de Machine Learning.....	403
4.2.	Preparación de los datos.....	403
4.2.1.	Librerías requeridas	403
4.2.2.	Carga de los conjuntos de datos.....	404
4.2.3.	Revisión inicial de los conjuntos de datos	408
4.2.4.	Limpieza inicial de datos	412
4.2.5.	Integración de los conjuntos de datos	415
4.2.6.	Vista previa del conjunto de datos unificado	461
4.2.7.	Limpieza de datos del conjunto de datos unificado	461
4.3.	Desarrollo de modelos de machine learning.....	494
4.3.1.	Librerías requeridas	494
4.3.2.	Selección de los modelos o algoritmos.....	495
4.3.3.	Entrenamiento y validación de los modelos	496
4.4.	Evaluación de los modelos de machine learning	518
CAPÍTULO V: RESULTADO		520
5.	Resultado.....	520
5.1.	Descripción de los conjuntos de datos iniciales.....	520

5.2.Resultados de la exploración de datos	521
5.3.Resultados del modelado del rendimiento académico	523
5.4.Comparación de modelos.....	524
CAPÍTULO VI: DISCUSIÓN	526
6. Discusión.....	526
6.1.Interpretación de resultados	526
6.1.1. Una posible relación entre la modalidad virtual, el descenso del rendimiento académico y el decrecimiento de la cantidad de estudiantes matriculados	526
6.1.2. Historia, becas socioeconómicas y sus beneficiarios.....	527
6.2.Contribuciones y limitaciones del estudio	528
6.3.Implicaciones para investigación futura	529
CAPÍTULO VII: DESPLIEGUE DEL MODELO	530
7. Despliegue del modelo.....	530
7.1.Despliegue del modelo de machine learning escogido	530
7.2.Resultados y evaluación del despliegue.....	558
CAPÍTULO VIII: CONCLUSIONES	563
8. Conclusiones	563
8.1.Conclusión general.....	563
8.2.Conclusiones específicas	563
8.3.Recomendaciones	564

BIBLIOGRFÍA	565
ANEXOS	574

ÍNDICE DE FIGURAS, GRÁFICOS Y TABLAS

ÍNDICE DE FIGURAS

Ilustración 1 Librerías requeridas para la fase de comprensión de los datos	27
Ilustración 2 Nomenclatura de los marcos de datos de los conjuntos de datos obtenidos	29
Ilustración 3 Función para la importación de conjuntos de datos y colocación de nombres dada una nomenclatura específica.....	30
Ilustración 4 Código para la importación de los datos en marcos de datos para su uso en el código.....	31
Ilustración 5 Resultado de la importación de los conjuntos de datos.....	32
Ilustración 6 Código para la impresión de las dimensiones de los conjuntos de datos importados.....	33
Ilustración 7 Dimensiones de los conjuntos de datos del reporte de Asignaturas Perdidas	34
Ilustración 8 Dimensiones de los conjuntos de datos del reporte de Calificaciones por Carrera.....	35
Ilustración 9 Dimensiones de los conjuntos de datos del reporte de Becas	36
Ilustración 10 Funciones generales para el análisis de la consistencia entre periodos de un reporte	37
Ilustración 11 Código para el análisis de consistencia del reporte de Asignaturas Perdidas.....	38

Ilustración 12 Parte 1 de los resultado del código para el análisis de las variables y su cantidad del reporte de Asignaturas Perdidas (corresponden al primer y segundo periodo del reporte).....	38
Ilustración 13 Parte 2 de los resultado del código para el análisis de las variables y su cantidad del reporte de Asignaturas Perdidas (corresponden al tercer y cuarto periodo del reporte).....	39
Ilustración 14 Parte Final de los resultado del código para el análisis de las variables y su cantidad del reporte de Asignaturas Perdidas (Resultado de la consistencia de las variables).....	39
Ilustración 15 Código para el análisis de consistencia del reporte de Calificaciones por Carrera.....	40
Ilustración 16 Parte 1 de los resultado del código para el análisis de las variables y su cantidad del reporte de Calificaciones por Carrera (corresponden al primer y segundo periodo del reporte).....	40
Ilustración 17 Parte 2 de los resultado del código para el análisis de las variables y su cantidad del reporte de Calificaciones por Carrera (corresponden al tercer y cuarto periodo del reporte).....	41
Ilustración 18 Parte Final de los resultado del código para el análisis de las variables y su cantidad del reporte de Calificaciones por Carrera (Resultado de la consistencia de las variables).....	41
Ilustración 19 Código para el análisis de consistencia de los periodos del reporte de Becas	42
Ilustración 20 Parte 2 de los resultado del código para el análisis de las variables y su cantidad del reporte de Becas (corresponden al primer y segundo periodo del reporte).....	42

Ilustración 21 Parte 2 de los resultado del código para el análisis de las variables y su cantidad del reporte de Becas (corresponden al tercer y cuarto periodo del reporte).....	43
Ilustración 22 Parte Final de los resultado del código para el análisis de las variables y su cantidad del reporte de Becas (Resultado de la consistencia de las variables).....	43
Ilustración 23 Visualización de la primera parte del primer periodo del reporte de Asignaturas Perdidas Anonimizado.....	62
Ilustración 24 Visualización de la parte final del primer periodo del reporte de Asignaturas Perdidas Anonimizado.....	63
Ilustración 25 Visualización de la primera parte del primer periodo del reporte de Calificaciones por Carrera Anonimizado.....	63
Ilustración 26 Visualización de la segunda parte del primer periodo del reporte de Calificaciones por Carrera Anonimizado.....	64
Ilustración 27 Visualización de la tercera parte del primer periodo del reporte de Calificaciones por Carrera Anonimizado.....	64
Ilustración 28 Visualización de la tercera parte del primer periodo del reporte de Calificaciones por Carrera Anonimizado.....	64
Ilustración 29 Visualización de la parte final del primer periodo del reporte de Calificaciones por Carrera Anonimizado.....	65
Ilustración 30 Visualización de la primera parte del primer periodo del reporte de Becas Anonimizado.....	65
Ilustración 31 Visualización de la segunda parte del primer periodo del reporte de Becas Anonimizado.....	65

Ilustración 32 Visualización de la parte final del primer periodo del reporte de Becas Anonimizado	66
Ilustración 33 Recapitulación de variables del reporte de Asignaturas Perdidas	68
Ilustración 34 Recapitulación de las variables del reporte de Calificaciones por Carrera	69
Ilustración 35 Recapitulación de las variables del reporte de Becas	70
Ilustración 36 Código utilizado para el análisis del contenido de las variables del reporte de Asignaturas Perdidas	71
Ilustración 37 Código utilizado para el análisis del contenido de las variables del reporte de Calificaciones por Carrera	71
Ilustración 38 Código utilizado para el análisis del contenido de las variables del reporte de Becas	72
Ilustración 39 Código para la comparativa de tipos de variables del reporte de Asignaturas Perdidas	81
Ilustración 40 Código para la comparativo de los tipos de variables del reporte de Calificaciones por Carrera	82
Ilustración 41 Código para la comparativa de los tipos de variables del reporte de Becas	83
Ilustración 42 Ejecución de la comparativa del reporte de Asignaturas Perdidas	84
Ilustración 43 Resultados de la comparativa del primer periodo del reporte de Asignaturas Perdidas	84
Ilustración 44 Resultados de la comparativa del segundo periodo del reporte de Asignaturas Perdidas	85

Ilustración 45 Resultados de la comparativo del tercer periodo del reporte de Asignaturas Perdidas.....	85
Ilustración 46 Resultados de la comparativa del cuarto periodo del reporte de Asignaturas Perdidas.....	86
Ilustración 47 Ejecución de la comparativa del reporte de Calificaciones por Carrera ..	86
Ilustración 48 Resultados de la comparativa del primer periodo del reporte de Calificaciones por Carrera	87
Ilustración 49 Resultados de la comparativa del segundo periodo del reporte de Calificaciones por Carrera	88
Ilustración 50 Resultados de la comparativa del tercer periodo del reporte de Calificaciones por Carrera	89
Ilustración 51 Resultados de la comparativa del cuarto periodo del reporte de Calificaciones por Carrera	90
Ilustración 52 Ejecución de la comparativa del reporte de Becas.....	90
Ilustración 53 Resultados de la comparativa del primer periodo del reporte de Becas...	91
Ilustración 54 Resultados de la comparativa del segundo periodo del reporte de Becas	91
Ilustración 55 Resultados de la comparativa del tercer periodo del reporte de Becas	92
Ilustración 56 Resultados de la comparativa del cuarto periodo del reporte de Becas ...	92
Ilustración 57 Función de ajuste de las variables del reporte de Asignaturas Perdidas ..	95
Ilustración 58 Función de ajuste de las variables del reporte de Calificaciones por Carrera.....	95
Ilustración 59 Función de ajuste de las variables del reporte de Becas	96
Ilustración 60 Aplicación del ajuste de las variables de los reportes	96

Ilustración 61 Resultados de la comparativa tras el ajuste del primer periodo del reporte de Asignaturas Perdidas	97
Ilustración 62 Resultados de la comparativa tras el ajuste del segundo periodo del reporte de Asignaturas Perdidas.....	97
Ilustración 63 Resultados de la comparativa tras el ajuste del tercer periodo del reporte de Asignaturas Perdidas	98
Ilustración 64 Resultados de la comparativa tras el ajuste del cuarto periodo del reporte de Asignaturas Perdidas	98
Ilustración 65 Resultados de la comparativa tras el ajuste del primer periodo del reporte de Calificaciones por Carrera.....	99
Ilustración 66 Resultados de la comparativa tras el ajuste del segundo periodo del reporte de Calificaciones por Carrera	100
Ilustración 67 Resultados de la comparativa tras el ajuste del tercer periodo del reporte de Calificaciones por Carrera.....	101
Ilustración 68 Resultados de la comparativa tras el ajuste del cuarto periodo del reporte de Calificaciones por Carrera.....	102
Ilustración 69 Resultados de la comparativa tras el ajuste del primer periodo del reporte de Becas	103
Ilustración 70 Resultados de la comparativa tras el ajuste del segundo periodo del reporte de Becas	104
Ilustración 71 Resultados de la comparativa tras el ajuste del tercer periodo del reporte de Becas	105

Ilustración 72 Resultados de la comparativa tras el ajuste del cuarto periodo del reporte de Becas	106
Ilustración 73 Código para la identificación de variables numéricas y categóricas del reporte de Asignaturas Perdidas.....	107
Ilustración 74 Tipos de variables identificadas del primer periodo del reporte de Asignaturas Perdidas.....	108
Ilustración 75 Tipos de variables identificadas del segundo periodo del reporte de Asignaturas Perdidas.....	109
Ilustración 76 Tipos de variables identificadas del tercer periodo del reporte de Asignaturas Perdidas.....	110
Ilustración 77 Tipos de variables identificadas del cuarto periodo del reporte de Asignaturas Perdidas.....	111
Ilustración 78 Código para la identificación de variables numéricas y categóricas del reporte de Calificaciones por Carrera	112
Ilustración 79 Tipos de variables identificadas del primer periodo del reporte de Calificaciones por Carrera	113
Ilustración 80 Tipos de variables identificadas del segundo periodo del reporte de Calificaciones por Carrera	114
Ilustración 81 Tipos de variables identificadas del tercer periodo del reporte de Calificaciones por Carrera	115
Ilustración 82 Tipos de variables identificadas del cuarto periodo del reporte de Calificaciones por Carrera	116

Ilustración 83 Código para la identificación de variables numéricas y categóricas del reporte de Becas	117
Ilustración 84 Tipos de variables identificadas del primero periodo del reporte de Becas	118
Ilustración 85 Tipos de variables identificadas del segundo periodo del reporte de Becas	119
Ilustración 86 Tipos de variables identificadas del tercer periodo del reporte de Becas	120
Ilustración 87 Tipos de variables identificadas del cuarto periodo del reporte de Becas	121
Ilustración 88 Código de eliminación de variables irrelevantes del reporte de Asignaturas Perdidas	130
Ilustración 89 Código de la eliminación de variables irrelevantes del reporte de Calificaciones por Carrera	131
Ilustración 90 Código de la eliminación y ajuste de variables irrelevantes del reporte de Becas	131
Ilustración 91 Vista del marco de datos del reporte de Asignaturas Perdidas	132
Ilustración 92 Vista del marco de datos del reporte de Calificaciones por Carrera	132
Ilustración 93 Parte 1 de la vista del marco de datos del reporte de Becas.....	132
Ilustración 94 Parte 2 de la vista del marco de datos del reporte de Becas.....	133
Ilustración 95 Identificación en código de tipos de variables relevantes del reporte de Asignaturas Perdidas.....	134

Ilustración 96 Identificación en código de tipos de variables relevantes del reporte de Calificaciones por Carrera	134
Ilustración 97 Identificación en código de tipos de variables relevantes del reporte de Becas	135
Ilustración 98 Estadísticas descriptivas del primer periodo del reporte de Asignaturas Perdidas	137
Ilustración 99 Estadísticas descriptivas del segundo periodo del reporte de Asignaturas Perdidas	139
Ilustración 100 Estadísticas descriptivas del tercer periodo del reporte de Asignaturas Perdidas	142
Ilustración 101 Estadísticas descriptivas del cuarto periodo del reporte de Asignaturas Perdidas	144
Ilustración 102 Código de la generación de un histograma de las notas finales del reporte de Asignaturas Perdidas	148
Ilustración 103 Histogramas del primer y segundo periodo de la variable NOTA_FINAL del reporte de asignaturas perdidas	149
Ilustración 104 Histogramas del tercer y cuarto periodo de la variable NOTA_FINAL del reporte de asignaturas perdidas	149
Ilustración 105 Código de la generación de un histograma de las repeticiones de un estudiante en una determinada asignatura del reporte de Asignaturas Perdidas	151
Ilustración 106 Histogramas del primer y segundo periodo de la variable VECES del reporte de Asignaturas Perdidas	151

Ilustración 107 Histogramas del tercer y cuarto periodo de la variable VECES del reporte de Asignaturas Perdidas.....	152
Ilustración 108 Código de la generación de un histograma de los créditos por asignatura del reporte de Asignaturas Perdidas.....	153
Ilustración 109 Histograma del primer periodo de la variable CRED del reporte de Asignaturas Perdidas.....	154
Ilustración 110 Histograma del segundo periodo de la variable CRED del reporte de Asignaturas Perdidas.....	155
Ilustración 111 Histograma del tercer periodo de la variable CRED del reporte de Asignaturas Perdidas.....	156
Ilustración 112 Histograma del cuarto periodo de la variable CRED del reporte de Asignaturas Perdidas.....	157
Ilustración 113 Código que genera el gráfico de densidad de las notas finales del reporte de Asignaturas Perdidas.....	158
Ilustración 114 Gráfico de densidad de la variable NOTA_FINAL del reporte de Asignaturas Finales.....	159
Ilustración 115 Código que genera el gráfico de densidad de la cantidad de veces que un estudiante repitió una misma asignatura del reporte de Asignaturas Perdidas	160
Ilustración 116 Gráfico de densidad de la variable VECES del reporte de Asignaturas Perdidas.....	161
Ilustración 117 Código que genera el gráfico de densidad de la cantidad de créditos de asignaturas del reporte de Asignaturas Perdidas.....	162

Ilustración 118 Gráfico de densidad de la variable CRED del reporte de Asignaturas Perdidas	163
Ilustración 119 Código para la generación de diagramas de caja y bigotes de las notas finales del reporte de Asignaturas Perdidas	164
Ilustración 120 Diagramas de caja y bigotes del primer y segundo periodo de la variable NOTA_FINAL del reporte de Asignaturas Perdidas.....	165
Ilustración 121 Diagramas de caja y bigotes del tercer y cuarto periodo de la variable NOT_FINAL del reporte de Asignaturas Perdidas.....	165
Ilustración 122 Estadísticas descriptivas del primer periodo del reporte de Calificaciones por Carrera	167
Ilustración 123 Estadísticas descriptivas del segundo periodo del reporte de Calificaciones por Carrera	169
Ilustración 124 Estadísticas descriptivas del tercer periodo del reporte de Calificaciones por Carrera	170
Ilustración 125 Estadísticas descriptivas del cuarto periodo del reporte de Calificaciones por Carrera	172
Ilustración 126 Código para la generación de los histogramas de las notas parciales del reporte de Calificaciones por Carrera	175
Ilustración 127 Histogramas del primer y segundo periodo de la variable nota_parcial del reporte de Calificaciones por Carrera	176
Ilustración 128 Histogramas del tercer y cuarto periodo de la variable nota_parcial del reporte de Calificaciones por Carrera	176

Ilustración 129 Código que genera los histogramas de las notas finales del reporte de Calificaciones por Carrera	177
Ilustración 130 Histogramas del primer y segundo periodo de la variable nota_final del reporte de Calificaciones por Carrera	178
Ilustración 131 Histogramas del tercer y cuarto periodo de la variable nota_final del reporte de Calificaciones por Carrera	178
Ilustración 132 Código que genera el gráfico de densidad de las notas parciales del reporte de Calificaciones por Carrera	180
Ilustración 133 Gráfico de densidad de la variable nota_parcial del reporte de Calificaciones por Carrera	181
Ilustración 134 Código que genera el gráfico de densidad de las notas finales del reporte de Calificaciones por Carrera.....	182
Ilustración 135 Gráfico de densidad de la variable nota_final del reporte de Calificaciones por Carrera	183
Ilustración 136 Código que genera los diagramas de caja y bigotes de las notas parciales del reporte de Calificaciones por Carrera	184
Ilustración 137 Diagrama de caja y bigotes del primer y segundo periodo de la variable nota_parcial del reporte de Calificaciones por Carrera.....	185
Ilustración 138 Diagrama de caja y bigotes del tercer y cuarto periodo de la variable nota_parcial del reporte de Calificaciones por Carrera.....	185
Ilustración 139 Código que genera los diagramas de caja y bigotes de las calificaciones finales del reporte de Calificaciones por Carrera.....	186

Ilustración 140 Diagramas de caja y bigotes del primer y segundo periodo de la variable nota_final del reporte de Calificaciones por Carrera	187
Ilustración 141 Diagramas de caja y bigotes del tercer y cuarto periodo de la variable nota_final del reporte de Calificaciones por Carrera	187
Ilustración 142 Estadísticas Descriptivas del primer periodo del reporte de Becas.....	189
Ilustración 143 Estadísticas Descriptivas del segundo periodo del reporte de Becas ...	190
Ilustración 144 Estadísticas Descriptivas del tercer periodo del reporte de Becas	191
Ilustración 145 Estadísticas Descriptivas del cuarto periodo del reporte de Becas	192
Ilustración 146 Código que genera los histogramas de la cantidad de segundas matrículas que poseen los estudiantes del reporte de Becas	194
Ilustración 147 Histogramas del primer y segundo periodo de la variable NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas	195
Ilustración 148 Histogramas del tercer y cuarto periodo de la variable NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas	195
Ilustración 149 Código que genera los histogramas de la cantidad de terceras matrículas que poseen los estudiantes del reporte de Becas.....	197
Ilustración 150 Histogramas del primer y segundo periodo de la variable NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas	197
Ilustración 151 Histogramas del tercer y cuarto periodo de la variable NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas	198
Ilustración 152 Código que genera el gráfico de densidad de la cantidad de segundas matrículas que poseen los estudiantes en el reporte de Becas	199

Ilustración 153 Gráfico de densidad de la variable	
NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas	200
Ilustración 154 Código para generar los gráficos de calor de la cantidad de terceras	
matrículas que poseen los estudiantes del reporte de Becas	201
Ilustración 155 Gráfico de densidad de la variable	
NUM_MATERIAS_TERCERA_MATIRCULA del reporte de Becas	202
Ilustración 156 Estadísticas descriptivas del primer periodo del reporte de Asignaturas	
Perdidas.....	205
Ilustración 157 Estadísticas descriptivas del segundo periodo del reporte de Asignaturas	
Perdidas.....	207
Ilustración 158 Estadísticas descriptivas del tercer periodo del reporte de Asignaturas	
Perdidas.....	209
Ilustración 159 Estadísticas descriptivas del cuarto periodo del reporte de Asignaturas	
Perdidas.....	211
Ilustración 160 Función necesaria para mejorar la visualización de los gráficos	214
Ilustración 161 Código que genera el gráfico de la variable PLAN_E del reporte de	
Asignaturas Perdidas.....	215
Ilustración 162 Gráficos de barras del primer y segundo periodo de la variable PLAN_E	
del reporte de Asignaturas Perdidas.....	215
Ilustración 163 Gráficos de barras del segundo y tercer periodo de la variable PLAN_E	
del reporte de Asignaturas Perdidas.....	216
Ilustración 164 Código que genera el gráfico de la variable DESC_FACULTAD del	
reporte de Asignaturas Perdidas.....	217

Ilustración 165 Gráficos de barras del primer y segundo periodo de la variable DESC_FACULTAD del reporte de Asignaturas Perdidas	218
Ilustración 166 Gráficos de barras del segundo y tercer periodo de la variable DESC_FACULTAD del reporte de Asignaturas Perdidas	219
Ilustración 167 Código que genera el gráfico de la variable DESC_PROGRAMA del reporte de Asignaturas Perdidas.....	221
Ilustración 168 Gráficos de barras del primer periodo de la variable DESC_PROGRAMA del reporte de Asignaturas Perdidas.....	222
Ilustración 169 Gráficos de barras del segundo periodo de la variable DESC_PROGRAMA del reporte de Asignaturas Perdidas.....	222
Ilustración 170 Gráficos de barras del tercer periodo de la variable DESC_PROGRAMA del reporte de Asignaturas Perdidas.....	223
Ilustración 171 Gráficos de barras del cuarto periodo de la variable DESC_PROGRAMA del reporte de Asignaturas Perdidas.....	223
Ilustración 172 Código que genera el gráfico de la variable NIVEL_ESTUDIANTE del reporte de Asignaturas Perdidas.....	225
Ilustración 173 Gráficos de barras del primer y segundo periodo de la variable NIVEL_ESTUDIANTE del reporte de Asignaturas Perdidas.....	225
Ilustración 174 Gráficos de barras del tercer y cuarto periodo de la variable NIVEL_ESTUDIANTE del reporte de Asignaturas Perdidas.....	226
Ilustración 175 Código que genera el gráfico de la variable MATE del reporte de Asignaturas Perdidas.....	227

Ilustración 176 Gráficos de barras del primer y segundo periodo de la variable MATE del reporte de Asignaturas Perdidas.....	228
Ilustración 177 Gráficos de barras del tercer y cuarto periodo de la variable MATE del reporte de Asignaturas Perdidas.....	228
Ilustración 178 Código que genera el gráfico de la variable APROBADO del reporte de Asignaturas Perdidas.....	230
Ilustración 179 Gráficos de pastel del primer y segundo periodo de la variable APROBADO del reporte de Asignaturas Perdidas.....	231
Ilustración 180 Gráficos de pastel del tercer y cuarto periodo de la variable APROBADO del reporte de Asignaturas Perdidas.....	231
Ilustración 181 Estadísticas Descriptivas del primer periodo del reporte de Calificaciones por Carrera	232
Ilustración 182 Estadísticas Descriptivas del segundo periodo del reporte de Calificaciones por Carrera	234
Ilustración 183 Estadísticas Descriptivas del tercer periodo del reporte de Calificaciones por Carrera	235
Ilustración 184 Estadísticas Descriptivas del cuarto periodo del reporte de Calificaciones por Carrera	237
Ilustración 185 Función necesaria para la mejor visualización de los gráficos	240
Ilustración 186 Código que genera los gráficos para la variable facultad del reporte de Calificaciones por Carrera	241
Ilustración 187 Gráfico de barras del primer y segundo periodo de la variable facultad del reporte de Calificaciones por Carrera	241

Ilustración 188 Gráfico de barras del tercer y cuarto periodo de la variable facultad del reporte de Calificaciones por Carrera	242
Ilustración 189 Código que genera los gráficos para la variable descripcion_programa del reporte de Calificaciones por Carrera	243
Ilustración 190 Gráfico de barras del primer periodo de la variable descripcion_programa del reporte de Calificaciones por Carrera.....	244
Ilustración 191 Gráfico de barras del segundo periodo de la variable descripcion_programa del reporte de Calificaciones por Carrera.....	244
Ilustración 192 Gráfico de barras del tercer periodo de la variable descripcion_programa del reporte de Calificaciones por Carrera	244
Ilustración 193 Gráfico de barras del cuarto periodo de la variable descripcion_programa del reporte de Calificaciones por Carrera	245
Ilustración 194 Código que genera los gráficos para la variable nota_final_SHATCKG del reporte de Calificaciones por Carrera	246
Ilustración 195 Gráfico de barras del primer periodo de la variable nota_final_SHATCKG del reporte de Calificaciones por Carrera	246
Ilustración 196 Gráfico de barras del segundo periodo de la variable nota_final_SHATCKG del reporte de Calificaciones por Carrera	247
Ilustración 197 Gráfico de barras del tercer periodo de la variable nota_final_SHATCKG del reporte de Calificaciones por Carrera	247
Ilustración 198 Gráfico de barras del cuarto periodo de la variable nota_final_SHATCKG del reporte de Calificaciones por Carrera	247

Ilustración 199 Parte 1 de las Estadísticas Descriptivas del primer periodo del reporte de Calificaciones por Carrera	249
Ilustración 200 Parte final de las Estadísticas Descriptivas del primer periodo del reporte de Calificaciones por Carrera.....	249
Ilustración 201 Parte 1 de las Estadísticas Descriptivas del segundo periodo del reporte de Calificaciones por Carrera.....	251
Ilustración 202 Parte final de las Estadísticas Descriptivas del segundo periodo del reporte de Calificaciones por Carrera	251
Ilustración 203 Parte 1 de las Estadísticas Descriptivas del tercer periodo del reporte de Calificaciones por Carrera	253
Ilustración 204 Parte final de las Estadísticas Descriptivas del tercer periodo del reporte de Calificaciones por Carrera.....	254
Ilustración 205 Parte 1 de las Estadísticas Descriptivas del cuarto periodo del reporte de Calificaciones por Carrera	256
Ilustración 206 Parte final de las Estadísticas Descriptivas del cuarto periodo del reporte de Calificaciones por Carrera.....	256
Ilustración 207 Función necesaria para la visualización coherente de los gráficos de los reportes.....	260
Ilustración 208 Código que genera los gráficos para la variable NACIONALIDAD del reporte de Becas	260
Ilustración 209 Gráfico de barras del primer y segundo periodo de la variable NACIONALIDAD del reporte de Becas	261

Ilustración 210 Gráfico de barras del tercer y cuarto periodo de la variable NACIONALIDAD del reporte de Becas	261
Ilustración 211 Código que genera los gráficos para la variable GENERO del reporte de Becas	262
Ilustración 212 Gráfico de pastel del primer y segundo periodo de la variable GENERO del reporte de Becas	263
Ilustración 213 Gráfico de pastel del tercer y cuarto periodo de la variable GENERO del reporte de Becas	263
Ilustración 214 Código que genera los gráficos para la variable ETNIA del reporte de Becas	264
Ilustración 215 Gráfico de barras del primer y segundo periodo de la variable ETNIA del reporte de Becas	264
Ilustración 216 Gráfico de barras del tercer y cuarto periodo de la variable ETNIA del reporte de Becas	265
Ilustración 217 Código que genera los gráficos para la variable TIPO_COLEGIO del reporte de Becas	266
Ilustración 218 Gráfico de barras del primer y segundo periodo de la variable TIPO_COLEGIO del reporte de Becas.....	266
Ilustración 219 Gráfico de barras del tercer y cuarto periodo de la variable TIPO_COLEGIO del reporte de Becas.....	267
Ilustración 220 Código que genera los gráficos para la variable PROGAMA_DESC del reporte de Becas	268

Ilustración 221 Gráfico de barras del primer periodo de la variable PROGAMA_DESC del reporte de Becas	268
Ilustración 222 Gráfico de barras del segundo periodo de la variable PROGAMA_DESC del reporte de Becas	269
Ilustración 223 Gráfico de barras del tercer periodo de la variable PROGAMA_DESC del reporte de Becas	269
Ilustración 224 Gráfico de barras del cuarto periodo de la variable PROGAMA_DESC del reporte de Becas	269
Ilustración 225 Código que genera los gráficos para la variable NIVEL_POR_MATERIAS del reporte de Becas.....	270
Ilustración 226 Gráfico de barras del primer y segundo periodo de la variable NIVEL_POR_MATERIAS del reporte de Becas.....	271
Ilustración 227 Gráfico de barras del tercer y cuarto periodo de la variable NIVEL_POR_MATERIAS del reporte de Becas.....	271
Ilustración 228 Código que genera los gráficos para la variable NIVEL_POR_ASIG_APROBADAS del reporte de Becas.....	272
Ilustración 229 Gráfico de barras del primer y segundo periodo de la variable NIVEL_POR_ASIG_APROBADAS del reporte de Becas.....	273
Ilustración 230 Gráfico de barras del tercer y cuarto periodo de la variable NIVEL_POR_ASIG_APROBADAS del reporte de Becas.....	273
Ilustración 231 Código que genera los gráficos para la variable CATEGORIA_BECA_SOCIOECONOMICA del reporte de Becas	274

Ilustración 232 Gráfico de barras del primer y segundo periodo de la variable	
CATEGORIA_BECA_SOCIOECONOMICA del reporte de Becas	274
Ilustración 233 Gráfico de barras del tercer y cuarto periodo de la variable	
CATEGORIA_BECA_SOCIOECONOMICA del reporte de Becas	275
Ilustración 234 Código que genera los gráficos para la variable TIENE_PAGO del	
reporte de Becas	276
Ilustración 235 Gráfico de barras del primer y segundo periodo de la variable	
TIENE_PAGO del reporte de Becas.....	276
Ilustración 236 Gráfico de barras del tercer y cuarto periodo de la variable	
TIENE_PAGO del reporte de Becas.....	277
Ilustración 237 Código que genera los gráficos para la variable TIENE_SALDO del	
reporte de Becas	278
Ilustración 238 Gráfico de barras del primer y segundo periodo de la variable	
TIENE_SALDO del reporte de Becas	278
Ilustración 239 Gráfico de barras del tercer y cuarto periodo de la variable	
TIENE_SALDO del reporte de Becas	279
Ilustración 240 Código que genera los gráficos para la relación entre PLAN_E y	
NOTA_FINAL del reporte de Asignaturas Perdidas.....	284
Ilustración 241 Diagrama de Violín del primer periodo de la relación entre PLAN_E y	
NOTA_FINAL del reporte de Asignaturas Perdidas.....	284
Ilustración 242 Diagrama de Violín del segundo periodo de la relación entre PLAN_E y	
NOTA_FINAL del reporte de Asignaturas Perdidas.....	285

Ilustración 243 Diagrama de Violín del tercer periodo de la relación entre PLAN_E y NOTA_FINAL del reporte de Asignaturas Perdidas.....	285
Ilustración 244 Diagrama de Violín del cuarto periodo de la relación entre PLAN_E y NOTA_FINAL del reporte de Asignaturas Perdidas.....	286
Ilustración 245 Código que genera los gráficos para la relación entre DESC_FACULTAD y NOTA_FINAL del reporte de Asignaturas Perdidas	287
Ilustración 246 Diagrama de Violín del primer periodo de la relación entre DESC_FACULTAD y NOTA_FINAL del reporte de Asignaturas Perdidas	288
Ilustración 247 Diagrama de Violín del segundo periodo de la relación entre DESC_FACULTAD y NOTA_FINAL del reporte de Asignaturas Perdidas	288
Ilustración 248 Diagrama de Violín del tercer periodo de la relación entre DESC_FACULTAD y NOTA_FINAL del reporte de Asignaturas Perdidas	289
Ilustración 249 Diagrama de Violín del cuarto periodo de la relación entre DESC_FACULTAD y NOTA_FINAL del reporte de Asignaturas Perdidas	289
Ilustración 250 Código que genera los gráficos para la relación entre DESC_PROGRAMA y NOTA_FINAL del reporte de Asignaturas Perdidas.....	291
Ilustración 251 Diagrama de Violín del primer periodo de la relación entre DESC_PROGRAMA y NOTA_FINAL del reporte de Asignaturas Perdidas.....	291
Ilustración 252 Diagrama de Violín del segundo periodo de la relación entre DESC_PROGRAMA y NOTA_FINAL del reporte de Asignaturas Perdidas.....	292
Ilustración 253 Diagrama de Violín del tercer periodo de la relación entre DESC_PROGRAMA y NOTA_FINAL del reporte de Asignaturas Perdidas.....	292

Ilustración 254 Diagrama de Violín del cuarto periodo de la relación entre DESC_PROGRAMA y NOTA_FINAL del reporte de Asignaturas Perdidas	293
Ilustración 255 Código que genera los gráficos para la relación entre NIVEL_ESTUDIANTE y NOTA_FINAL del reporte de Asignaturas Perdidas	295
Ilustración 256 Diagrama de Violín del primer periodo de la relación entre NIVEL_ESTUDIANTE y NOTA_FINAL del reporte de Asignaturas Perdidas	296
Ilustración 257 Diagrama de Violín del segundo periodo de la relación entre NIVEL_ESTUDIANTE y NOTA_FINAL del reporte de Asignaturas Perdidas	296
Ilustración 258 Diagrama de Violín del tercer periodo de la relación entre NIVEL_ESTUDIANTE y NOTA_FINAL del reporte de Asignaturas Perdidas	297
Ilustración 259 Diagrama de Violín del cuarto periodo de la relación entre NIVEL_ESTUDIANTE y NOTA_FINAL del reporte de Asignaturas Perdidas	297
Ilustración 260 Código que genera los gráficos para la relación entre MATE y NOTA_FINAL del reporte de Asignaturas Perdidas	298
Ilustración 261 Diagrama de Violín del primer periodo de la relación entre MATE y NOTA_FINAL del reporte de Asignaturas Perdidas	298
Ilustración 262 Diagrama de Violín del segundo periodo de la relación entre MATE y NOTA_FINAL del reporte de Asignaturas Perdidas	299
Ilustración 263 Diagrama de Violín del tercer periodo de la relación entre MATE y NOTA_FINAL del reporte de Asignaturas Perdidas	299
Ilustración 264 Diagrama de Violín del cuarto periodo de la relación entre MATE y NOTA_FINAL del reporte de Asignaturas Perdidas	300

Ilustración 265 Código que genera los gráficos para la relación entre VECES y	
NOTA_FINAL del reporte de Asignaturas Perdidas.....	301
Ilustración 266 Diagrama de Violín del primer periodo de la relación entre VECES y	
NOTA_FINAL del reporte de Asignaturas Perdidas.....	301
Ilustración 267 Diagrama de Violín del segundo periodo de la relación entre VECES y	
NOTA_FINAL del reporte de Asignaturas Perdidas.....	302
Ilustración 268 Diagrama de Violín del tercer periodo de la relación entre VECES y	
NOTA_FINAL del reporte de Asignaturas Perdidas.....	302
Ilustración 269 Diagrama de Violín del cuarto periodo de la relación entre VECES y	
NOTA_FINAL del reporte de Asignaturas Perdidas.....	303
Ilustración 270 Código que genera los gráficos para la relación entre CRED y	
NOTA_FINAL del reporte de Asignaturas Perdidas.....	304
Ilustración 271 Diagrama de Violín del primer periodo de la relación entre CRED y	
NOTA_FINAL del reporte de Asignaturas Perdidas.....	304
Ilustración 272 Diagrama de Violín del segundo periodo de la relación entre CRED y	
NOTA_FINAL del reporte de Asignaturas Perdidas.....	305
Ilustración 273 Diagrama de Violín del tercer periodo de la relación entre CRED y	
NOTA_FINAL del reporte de Asignaturas Perdidas.....	305
Ilustración 274 Diagrama de Violín del cuarto periodo de la relación entre CRED y	
NOTA_FINAL del reporte de Asignaturas Perdidas.....	306
Ilustración 275 Código que genera los gráficos para la relación entre PLAN_E y VECES	
del reporte de Asignaturas Perdidas.....	307

Ilustración 276 Diagrama de Violín del primer periodo de la relación entre PLAN_E y VECES del reporte de Asignaturas Perdidas	307
Ilustración 277 Diagrama de Violín del segundo periodo de la relación entre PLAN_E y VECES del reporte de Asignaturas Perdidas	308
Ilustración 278 Diagrama de Violín del tercer periodo de la relación entre PLAN_E y VECES del reporte de Asignaturas Perdidas	308
Ilustración 279 Diagrama de Violín del cuarto periodo de la relación entre PLAN_E y VECES del reporte de Asignaturas Perdidas	309
Ilustración 280 Código que genera los gráficos para la relación entre DESC_FACULTAD y VECES del reporte de Asignaturas Perdidas	310
Ilustración 281 Diagrama de Violín del primer periodo de la relación entre DESC_FACULTAD y VECES del reporte de Asignaturas Perdidas	310
Ilustración 282 Diagrama de Violín del segundo periodo de la relación entre DESC_FACULTAD y VECES del reporte de Asignaturas Perdidas	311
Ilustración 283 Diagrama de Violín del tercer periodo de la relación entre DESC_FACULTAD y VECES del reporte de Asignaturas Perdidas	311
Ilustración 284 Diagrama de Violín del cuarto periodo de la relación entre DESC_FACULTAD y VECES del reporte de Asignaturas Perdidas	312
Ilustración 285 Código que genera los gráficos para la relación entre DESC_PROGRAMA y VECES del reporte de Asignaturas Perdidas	313
Ilustración 286 Diagrama de Violín del primer periodo de la relación entre DESC_PROGRAMA y VECES del reporte de Asignaturas Perdidas	313

Ilustración 287 Diagrama de Violín del segundo periodo de la relación entre DESC_PROGRAMA y VECES del reporte de Asignaturas Perdidas	314
Ilustración 288 Diagrama de Violín del tercer periodo de la relación entre DESC_PROGRAMA y VECES del reporte de Asignaturas Perdidas	314
Ilustración 289 Diagrama de Violín del cuarto periodo de la relación entre DESC_PROGRAMA y VECES del reporte de Asignaturas Perdidas	315
Ilustración 290 Código que genera los gráficos para la relación entre NIVEL_ESTUDIANTE y VECES del reporte de Asignaturas Perdidas	316
Ilustración 291 Diagrama de Violín del primer periodo de la relación entre NIVEL_ESTUDIANTE y VECES del reporte de Asignaturas Perdidas	316
Ilustración 292 Diagrama de Violín del segundo periodo de la relación entre NIVEL_ESTUDIANTE y VECES del reporte de Asignaturas Perdidas	317
Ilustración 293 Diagrama de Violín del tercer periodo de la relación entre NIVEL_ESTUDIANTE y VECES del reporte de Asignaturas Perdidas	317
Ilustración 294 Diagrama de Violín del cuarto periodo de la relación entre NIVEL_ESTUDIANTE y VECES del reporte de Asignaturas Perdidas	318
Ilustración 295 Código que genera los gráficos para la relación entre MATE y VECES del reporte de Asignaturas Perdidas.....	319
Ilustración 296 Diagrama de Violín del primer periodo de la relación entre MATE y VECES del reporte de Asignaturas Perdidas	319
Ilustración 297 Diagrama de Violín del segundo periodo de la relación entre MATE y VECES del reporte de Asignaturas Perdidas	320

Ilustración 298 Diagrama de Violín del tercer periodo de la relación entre MATE y VECES del reporte de Asignaturas Perdidas	320
Ilustración 299 Diagrama de Violín del cuarto periodo de la relación entre MATE y VECES del reporte de Asignaturas Perdidas	321
Ilustración 300 Código que genera los gráficos para la relación entre CRED y VECES del reporte de Asignaturas Perdidas.....	322
Ilustración 301 Diagrama de Violín del primer periodo de la relación entre CRED y VECES del reporte de Asignaturas Perdidas	322
Ilustración 302 Diagrama de Violín del segundo periodo de la relación entre CRED y VECES del reporte de Asignaturas Perdidas	323
Ilustración 303 Diagrama de Violín del tercer periodo de la relación entre CRED y VECES del reporte de Asignaturas Perdidas	323
Ilustración 304 Diagrama de Violín del cuarto periodo de la relación entre CRED y VECES del reporte de Asignaturas Perdidas	324
Ilustración 305 Código que genera los gráficos para las variables NOTA_FINAL, CRED y VECES del reporte de Asignaturas Perdidas	325
Ilustración 306 Mapa de calor del primer periodo de las variables NOTA_FINAL, CRED y VECES del reporte de Asignaturas Perdidas	325
Ilustración 307 Mapa de calor del segundo periodo de las variables NOTA_FINAL, CRED y VECES del reporte de Asignaturas Perdidas	326
Ilustración 308 Mapa de calor del tercer periodo de las variables NOTA_FINAL, CRED y VECES del reporte de Asignaturas Perdidas	326

Ilustración 309 Mapa de calor del cuarto periodo de las variables NOTA_FINAL, CRED y VECES del reporte de Asignaturas Perdidas	327
Ilustración 310 Código que genera los gráficos para la relación entre facultad y nota_final del reporte de Calificaciones por Carrera	330
Ilustración 311 Diagrama de Violín del primer periodo de la relación entre facultad y nota_final del reporte de Calificaciones por Carrera	330
Ilustración 312 Diagrama de Violín del segundo periodo de la relación entre facultad y nota_final del reporte de Calificaciones por Carrera	331
Ilustración 313 Diagrama de Violín del tercer periodo de la relación entre facultad y nota_final del reporte de Calificaciones por Carrera	331
Ilustración 314 Diagrama de Violín del cuarto periodo de la relación entre facultad y nota_final del reporte de Calificaciones por Carrera	332
Ilustración 315 Código que genera los gráficos del primer muestreo para la relación entre id_banner_docente y nota_final del reporte de Calificaciones por Carrera.....	333
Ilustración 316 Diagrama de Violín del primer muestreo del primer periodo de la relación entre id_banner_docente y nota_final del reporte de Calificaciones por Carrera.....	333
Ilustración 317 Diagrama de Violín del primer muestreo del segundo periodo de la relación entre id_banner_docente y nota_final del reporte de Calificaciones por Carrera.....	334
Ilustración 318 Diagrama de Violín del primer muestreo del tercer periodo de la relación entre id_banner_docente y nota_final del reporte de Calificaciones por Carrera.....	334
Ilustración 319 Diagrama de Violín del primer muestreo del cuarto periodo de la relación entre id_banner_docente y nota_final del reporte de Calificaciones por Carrera.....	334

Ilustración 320 Código que genera los gráficos del segundo muestreo para la relación entre id_banner_docente y nota_final del reporte de Calificaciones por Carrera.....	335
Ilustración 321 Diagrama de Violín del segundo muestreo del primer periodo de la relación entre id_banner_docente y nota_final del reporte de Calificaciones por Carrera.....	335
Ilustración 322 Diagrama de Violín del segundo muestreo del segundo periodo de la relación entre id_banner_docente y nota_final del reporte de Calificaciones por Carrera.....	336
Ilustración 323 Diagrama de Violín del segundo muestreo del tercer periodo de la relación entre id_banner_docente y nota_final del reporte de Calificaciones por Carrera.....	336
Ilustración 324 Diagrama de Violín del segundo muestreo del cuarto periodo de la relación entre id_banner_docente y nota_final del reporte de Calificaciones por Carrera.....	336
Ilustración 325 Código que genera los gráficos del tercer muestreo para la relación entre id_banner_docente y nota_final del reporte de Calificaciones por Carrera.....	337
Ilustración 326 Diagrama de Violín del tercer muestreo del primer periodo de la relación entre id_banner_docente y nota_final del reporte de Calificaciones por Carrera.....	337
Ilustración 327 Diagrama de Violín del tercer muestreo del segundo periodo de la relación entre id_banner_docente y nota_final del reporte de Calificaciones por Carrera.....	338
Ilustración 328 Diagrama de Violín del tercer muestreo del tercer periodo de la relación entre id_banner_docente y nota_final del reporte de Calificaciones por Carrera.....	338
Ilustración 329 Diagrama de Violín del tercer muestreo del cuarto periodo de la relación entre id_banner_docente y nota_final del reporte de Calificaciones por Carrera.....	338
Ilustración 330 Código que genera los gráficos para la relación entre descripcion_materia y nota_final del reporte de Calificaciones por Carrera.....	339

Ilustración 331 Diagrama de Violín del primer periodo de la relación entre descripcion_materia y nota_final del reporte de Calificaciones por Carrera.....	340
Ilustración 332 Diagrama de Violín del segundo periodo de la relación entre descripcion_materia y nota_final del reporte de Calificaciones por Carrera.....	340
Ilustración 333 Diagrama de Violín del tercer periodo de la relación entre descripcion_materia y nota_final del reporte de Calificaciones por Carrera.....	340
Ilustración 334 Diagrama de Violín del cuarto periodo de la relación entre descripcion_materia y nota_final del reporte de Calificaciones por Carrera.....	341
Ilustración 335 Código que genera los gráficos para la relación entre descripcion_programa y nota_final del reporte de Calificaciones por Carrera	342
Ilustración 336 Diagrama de Caja y Bigotes del primer periodo de la relación entre descripcion_programa y nota_final del reporte de Calificaciones por Carrera	342
Ilustración 337 Diagrama de Caja y Bigotes del segundo periodo de la relación entre descripcion_programa y nota_final del reporte de Calificaciones por Carrera	343
Ilustración 338 Diagrama de Caja y Bigotes del tercer periodo de la relación entre descripcion_programa y nota_final del reporte de Calificaciones por Carrera	343
Ilustración 339 Diagrama de Caja y Bigotes del cuarto periodo de la relación entre descripcion_programa y nota_final del reporte de Calificaciones por Carrera	343
Ilustración 340 Código que genera los gráficos para la relación entre GENERO y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas	346
Ilustración 341 Diagrama de Violín del primer periodo de la relación entre GENERO y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas	346

Ilustración 342 Diagrama de Violín del segundo periodo de la relación entre GENERO y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas 347

Ilustración 343 Diagrama de Violín del tercer periodo de la relación entre GENERO y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas 347

Ilustración 344 Diagrama de Violín del cuarto periodo de la relación entre GENERO y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas 347

Ilustración 345 Código que genera los gráficos para la relación entre ETNIA y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas 348

Ilustración 346 Diagrama de Violín del primer periodo de la relación entre ETNIA y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas 348

Ilustración 347 Diagrama de Violín del segundo periodo de la relación entre ETNIA y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas 349

Ilustración 348 Diagrama de Violín del tercer periodo de la relación entre ETNIA y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas 349

Ilustración 349 Diagrama de Violín del cuarto periodo de la relación entre ETNIA y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas 349

Ilustración 350 Código que genera los gráficos para la relación entre TIPO_COLEGIO y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas 350

Ilustración 351 Diagrama de Violín del primer periodo de la relación entre TIPO_COLEGIO y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas 351

Ilustración 352 Diagrama de Violín del segundo periodo de la relación entre TIPO_COLEGIO y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas 351

Ilustración 353 Diagrama de Violín del tercer periodo de la relación entre TIPO_COLEGIO y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas 351

Ilustración 354 Diagrama de Violín del cuarto periodo de la relación entre TIPO_COLEGIO y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas 352

Ilustración 355 Código que genera los gráficos para la relación entre NIVEL_POR_MATERIAS y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas 352

Ilustración 356 Diagrama de Violín del primer periodo de la relación entre NIVEL_POR_MATERIAS y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas 353

Ilustración 357 Diagrama de Violín del segundo periodo de la relación entre NIVEL_POR_MATERIAS y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas 353

Ilustración 358 Diagrama de Violín del tercer periodo de la relación entre NIVEL_POR_MATERIAS y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas 353

Ilustración 359 Diagrama de Violín del cuarto periodo de la relación entre NIVEL_POR_MATERIAS y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas 354

Ilustración 360 Código que genera los gráficos para la relación entre NIVEL_POR_ASIG_APROBADAS y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas 354

Ilustración 361 Diagrama de Violín del primer periodo de la relación entre NIVEL_POR_ASIG_APROBADAS y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas 355

Ilustración 362 Diagrama de Violín del segundo periodo de la relación entre NIVEL_POR_ASIG_APROBADAS y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas 355

Ilustración 363 Diagrama de Violín del tercer periodo de la relación entre NIVEL_POR_ASIG_APROBADAS y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas 355

Ilustración 364 Diagrama de Violín del cuarto periodo de la relación entre NIVEL_POR_ASIG_APROBADAS y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas 356

Ilustración 365 Código que genera los gráficos para la relación entre CATEGORIA_BECA_SOCIOECONOMICA y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas 356

Ilustración 366 Diagrama de Violín del primer periodo de la relación entre CATEGORIA_BECA_SOCIOECONOMICA y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas 356

Ilustración 367 Diagrama de Violín del tercer periodo de la relación entre CATEGORIA_BECA_SOCIOECONOMICA y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas 357

Ilustración 368 Código que genera los gráficos para la relación entre GENERO y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas 358

Ilustración 369 Diagrama de Violín del primer periodo de la relación entre GENERO y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas	358
Ilustración 370 Diagrama de Violín del segundo periodo de la relación entre GENERO y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas	359
Ilustración 371 Diagrama de Violín del tercer periodo de la relación entre GENERO y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas	359
Ilustración 372 Diagrama de Violín del cuarto periodo de la relación entre GENERO y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas	359
Ilustración 373 Código que genera los gráficos para la relación entre ETNIA y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas	360
Ilustración 374 Diagrama de Violín del primer periodo de la relación entre ETNIA y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas	360
Ilustración 375 Diagrama de Violín del segundo periodo de la relación entre ETNIA y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas	361
Ilustración 376 Diagrama de Violín del tercer periodo de la relación entre ETNIA y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas	361
Ilustración 377 Diagrama de Violín del cuarto periodo de la relación entre ETNIA y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas	361
Ilustración 378 Código que genera los gráficos para la relación entre TIPO_COLEGIO y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas	362
Ilustración 379 Diagrama de Violín del primer periodo de la relación entre TIPO_COLEGIO y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas.....	362

Ilustración 380 Diagrama de Violín del segundo periodo de la relación entre TIPO_COLEGIO y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas.....	363
Ilustración 381 Diagrama de Violín del tercer periodo de la relación entre TIPO_COLEGIO y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas.....	363
Ilustración 382 Diagrama de Violín del cuarto periodo de la relación entre TIPO_COLEGIO y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas.....	363
Ilustración 383 Código que genera los gráficos para la relación entre NIVEL_POR_MATERIAS y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas.....	364
Ilustración 384 Diagrama de Violín del primer periodo de la relación entre NIVEL_POR_MATERIAS y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas.....	364
Ilustración 385 Diagrama de Violín del segundo periodo de la relación entre NIVEL_POR_MATERIAS y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas.....	364
Ilustración 386 Diagrama de Violín del tercer periodo de la relación entre NIVEL_POR_MATERIAS y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas.....	365
Ilustración 387 Diagrama de Violín del cuarto periodo de la relación entre NIVEL_POR_MATERIAS y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas.....	365

Ilustración 388 Código que genera los gráficos para la relación entre NIVEL_POR_ASIG_APROBADAS y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas 366

Ilustración 389 Diagrama de Violín del primer periodo de la relación entre NIVEL_POR_ASIG_APROBADAS y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas 366

Ilustración 390 Diagrama de Violín del segundo periodo de la relación entre NIVEL_POR_ASIG_APROBADAS y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas 366

Ilustración 391 Diagrama de Violín del tercer periodo de la relación entre NIVEL_POR_ASIG_APROBADAS y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas 367

Ilustración 392 Diagrama de Violín del cuarto periodo de la relación entre NIVEL_POR_ASIG_APROBADAS y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas 367

Ilustración 393 Código que genera los gráficos para la relación entre CATEGORIA_BECA_SOCIOECONOMICA y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas 367

Ilustración 394 Diagrama de Violín del primer periodo de la relación entre CATEGORIA_BECA_SOCIOECONOMICA y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas 368

Ilustración 395 Diagrama de Violín del tercer periodo de la relación entre CATEGORIA_BECA_SOCIOECONOMICA y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas	368
Ilustración 396 Verificación actual de dimensiones de los periodos de becas.....	370
Ilustración 397 Realización de la unión en marcos de datos temporales	371
Ilustración 398 Verificación de los marcos de datos temporales con respecto a los marcos de datos originales	372
Ilustración 399 Parte 1 de la vista del marco de datos del primer periodo del reporte de Becas original.....	372
Ilustración 400 Parte final de la vista del marco de datos del primer periodo del reporte de Becas original.....	372
Ilustración 401 Parte 1 de la vista del marco de datos del primer periodo del reporte de Becas temporal.....	372
Ilustración 402 Parte final de la vista del marco de datos del primer periodo del reporte de Becas temporal	373
Ilustración 403 Integración de la unión en los datos oficiales	373
Ilustración 404 Parte 1 de la vista del marco de datos unificado del primer periodo del reporte de Becas	373
Ilustración 405 Parte final de la vista del marco de datos unificado del primer periodo del reporte de Becas	373
Ilustración 406 Código que genera los gráficos para la relación entre GENERO y NOTA_FINAL del reporte de Becas	374

Ilustración 407 Diagrama de Violín del primer periodo de la relación entre GENERO y NOTA_FINAL del reporte de Becas	375
Ilustración 408 Diagrama de Violín del segundo periodo de la relación entre GENERO y NOTA_FINAL del reporte de Becas	375
Ilustración 409 Diagrama de Violín del tercer periodo de la relación entre GENERO y NOTA_FINAL del reporte de Becas	375
Ilustración 410 Diagrama de Violín del cuarto periodo de la relación entre GENERO y NOTA_FINAL del reporte de Becas	376
Ilustración 411 Código que genera los gráficos para la relación entre ETNIA y NOTA_FINAL del reporte de Becas	377
Ilustración 412 Diagrama de Violín del primer periodo de la relación entre ETNIA y NOTA_FINAL del reporte de Becas	377
Ilustración 413 Diagrama de Violín del segundo periodo de la relación entre ETNIA y NOTA_FINAL del reporte de Becas	378
Ilustración 414 Diagrama de Violín del tercer periodo de la relación entre ETNIA y NOTA_FINAL del reporte de Becas	378
Ilustración 415 Diagrama de Violín del cuarto periodo de la relación entre ETNIA y NOTA_FINAL del reporte de Becas	378
Ilustración 416 Código que genera los gráficos para la relación entre TIPO_COLEGIO y NOTA_FINAL del reporte de Becas	379
Ilustración 417 Diagrama de Violín del primer periodo de la relación entre TIPO_COLEGIO y NOTA_FINAL del reporte de Becas.....	380

Ilustración 418 Diagrama de Violín del segundo periodo de la relación entre TIPO_COLEGIO y NOTA_FINAL del reporte de Becas.....	380
Ilustración 419 Diagrama de Violín del tercer periodo de la relación entre TIPO_COLEGIO y NOTA_FINAL del reporte de Becas.....	380
Ilustración 420 Diagrama de Violín del cuarto periodo de la relación entre TIPO_COLEGIO y NOTA_FINAL del reporte de Becas.....	381
Ilustración 421 Código que genera los gráficos para la relación entre NIVEL_POR_ASIG_APROBADAS y NOTA_FINAL del reporte de Becas.....	381
Ilustración 422 Diagrama de Violín del primer periodo de la relación entre NIVEL_POR_ASIG_APROBADAS y NOTA_FINAL del reporte de Becas.....	382
Ilustración 423 Diagrama de Violín del segundo periodo de la relación entre NIVEL_POR_ASIG_APROBADAS y NOTA_FINAL del reporte de Becas.....	382
Ilustración 424 Diagrama de Violín del tercer periodo de la relación entre NIVEL_POR_ASIG_APROBADAS y NOTA_FINAL del reporte de Becas.....	382
Ilustración 425 Diagrama de Violín del cuarto periodo de la relación entre NIVEL_POR_ASIG_APROBADAS y NOTA_FINAL del reporte de Becas.....	383
Ilustración 426 Código que genera los gráficos para la relación entre CATEGORIA_BECA_SOCIOECONOMICA y NOTA_FINAL del reporte de Becas	383
Ilustración 427 Diagrama de Violín del primer periodo de la relación entre CATEGORIA_BECA_SOCIOECONOMICA y NOTA_FINAL del reporte de Becas	384
Ilustración 428 Diagrama de Violín del tercer periodo de la relación entre CATEGORIA_BECA_SOCIOECONOMICA y NOTA_FINAL del reporte de Becas	384

Ilustración 429 Código que genera los gráficos para la relación entre NUM_MATERIAS_SEGUNDA_MATRICULA y NOTA_FINAL del reporte de Becas	385
Ilustración 430 Diagrama de Violín del primer periodo de la relación entre NUM_MATERIAS_SEGUNDA_MATRICULA y NOTA_FINAL del reporte de Becas	385
Ilustración 431 Diagrama de Violín del segundo periodo de la relación entre NUM_MATERIAS_SEGUNDA_MATRICULA y NOTA_FINAL del reporte de Becas	385
Ilustración 432 Diagrama de Violín del tercer periodo de la relación entre NUM_MATERIAS_SEGUNDA_MATRICULA y NOTA_FINAL del reporte de Becas	386
Ilustración 433 Diagrama de Violín del cuarto periodo de la relación entre NUM_MATERIAS_SEGUNDA_MATRICULA y NOTA_FINAL del reporte de Becas	386
Ilustración 434 Código que genera los gráficos para la relación entre NUM_MATERIAS_TERCERA_MATRICULA y NOTA_FINAL del reporte de Becas	387
Ilustración 435 Diagrama de Violín del primer periodo de la relación entre NUM_MATERIAS_TERCERA_MATRICULA y NOTA_FINAL del reporte de Becas	387
Ilustración 436 Diagrama de Violín del segundo periodo de la relación entre NUM_MATERIAS_TERCERA_MATRICULA y NOTA_FINAL del reporte de Becas	387
Ilustración 437 Diagrama de Violín del tercer periodo de la relación entre NUM_MATERIAS_TERCERA_MATRICULA y NOTA_FINAL del reporte de Becas	388
Ilustración 438 Diagrama de Violín del cuarto periodo de la relación entre NUM_MATERIAS_TERCERA_MATRICULA y NOTA_FINAL del reporte de Becas	388
Ilustración 439 Código de la eliminación y verificación de esta de las variables del reporte de Asignaturas Perdidas.....	393

Ilustración 440 Resultados del primer y segundo periodo de la eliminación de variables del reporte de Asignaturas Perdidas.....	394
Ilustración 441 Resultados del tercer y cuarto periodo de la eliminación de variables del reporte de Asignaturas Perdidas.....	394
Ilustración 442 Código de la eliminación y verificación de esta de las variables del reporte de Calificaciones por Carrera	395
Ilustración 443 Resultados del primer y segundo periodo de la eliminación de variables del reporte de Calificaciones por Carrera	395
Ilustración 444 Resultados del tercer y cuarto periodo de la eliminación de variables del reporte de Calificaciones por Carrera	396
Ilustración 445 Código de la eliminación y verificación de esta de las variables del reporte de Becas	396
Ilustración 446 Resultados del primer y segundo periodo de la eliminación de variables del reporte de Becas	397
Ilustración 447 Resultados del tercer y cuarto periodo de la eliminación de variables del reporte de Becas	397
Ilustración 448 Vista del primer periodo del reporte de Asignaturas Perdidas	398
Ilustración 449 Vista del segundo periodo del reporte de Asignaturas Perdidas	398
Ilustración 450 Vista del tercer periodo del reporte de Asignaturas Perdidas	398
Ilustración 451 Vista del cuarto periodo del reporte de Asignaturas Perdidas	399
Ilustración 452 Vista del primer periodo del reporte de Calificaciones por Carrera	399
Ilustración 453 Vista del segundo periodo del reporte de Calificaciones por Carrera..	399
Ilustración 454 Vista del tercer periodo del reporte de Calificaciones por Carrera	400

Ilustración 455 Vista del cuarto periodo del reporte de Calificaciones por Carrera	400
Ilustración 456 Vista del primer periodo del reporte de Becas	400
Ilustración 457 Vista del segundo periodo del reporte de Becas	401
Ilustración 458 Vista del tercer periodo del reporte de Becas.....	401
Ilustración 459 Vista del cuarto periodo del reporte de Becas.....	401
Ilustración 460 Código de exportación de los datos pre-preparados.....	402
Ilustración 461 Librerías requeridas para la fase de preparación de los datos	404
Ilustración 462 Nomenclatura de los marcos de datos de los conjuntos de datos obtenidos	405
Ilustración 463 Función para la importación de conjuntos de datos y colocación de nombres dada una nomenclatura específica.....	406
Ilustración 464 Código para la importación de los datos en marcos de datos para su uso en la fase de preparación de datos.....	407
Ilustración 465 Resultados de la importación de los datos	408
Ilustración 466 Vista del primer periodo del reporte de Asignaturas Perdidas.....	411
Ilustración 467 Vista del primer periodo del reporte de Calificaciones por Carrera	411
Ilustración 468 Vista del primer periodo del reporte de Becas	411
Ilustración 469 Código para la eliminación de duplicados del reporte de Asignaturas Perdidas.....	413
Ilustración 470 Código para la eliminación de duplicados del reporte de Calificaciones por Carrera	413
Ilustración 471 Código para la eliminación de duplicados del reporte de Calificaciones por Carrera	414

Ilustración 472 Código para la identificación de las variables de los reportes	418
Ilustración 473 Identificación de las variables de los reportes.....	419
Ilustración 474 Definición de diccionarios para el cambio de nombre de las variables de los reportes	421
Ilustración 475 Código de la ejecución de la armonización de los nombres de las variables	422
Ilustración 476 Identificación de variables de los reportes tras la armonización.....	422
Ilustración 477 Identificación de los nombres de las variables armonizadas.....	423
Ilustración 478 Comparativa de tipos de datos del primer periodo de los datos.....	425
Ilustración 479 Comparativa de tipos de datos del segundo periodo de los datos	426
Ilustración 480 Comparativa de tipos de datos del tercer periodo de los datos	427
Ilustración 481 Comparativa de tipos de datos del cuarto periodo de los datos	428
Ilustración 482 Código para llevar a cabo el análisis de la variable area_asignatura ...	430
Ilustración 483 Resultado del primer periodo del análisis de la variable area_asignatura	430
Ilustración 484 Resultado del segundo periodo del análisis de la variable area_asignatura	431
Ilustración 485 Resultado del tercer periodo del análisis de la variable area_asignatura	432
Ilustración 486 Resultado del cuarto periodo del análisis de la variable area_asignatura	433
Ilustración 487 Código para la armonización de la variable area_asignatura	434
Ilustración 488 Vista previa de la armonización de la variable area_asignatura	434

Ilustración 489 Código para llevar a cabo el análisis de la variable nivel	435
Ilustración 490 Resultado del primer periodo del análisis de la variable nivel.....	435
Ilustración 491 Resultado del segundo periodo del análisis de la variable nivel	436
Ilustración 492 Resultado del tercer periodo del análisis de la variable nivel	437
Ilustración 493 Resultado del cuarto periodo del análisis de la variable nivel	438
Ilustración 494 Código de armonización de la variable nivel del cuarto periodo del reporte de Asignaturas Perdidas.....	439
Ilustración 495 Vista previa de la armonización de la variable nivel	439
Ilustración 496 Análisis de la compatibilidad entre variables armonizadas	440
Ilustración 497 Resultado del primer periodo del análisis de compatibilidad	440
Ilustración 498 Resultado del segundo periodo del análisis de compatibilidad.....	441
Ilustración 499 Resultado del tercer periodo del análisis de compatibilidad	441
Ilustración 500 Resultado del cuarto periodo del análisis de compatibilidad	441
Ilustración 501 Unión del primer periodo reporte de Calificaciones por Carrera y el reporte de Becas	443
Ilustración 502 Evaluación del estado de la unión	443
Ilustración 503 Evaluación específica de la unión	443
Ilustración 504 Unión del primer periodo del reporte de asignaturas perdidas con la unión inicial.....	444
Ilustración 505 Evaluación del estado de la unión completa	444
Ilustración 506 Evaluación específica de la unión completa	444
Ilustración 507 Unión del segundo periodo reporte de Calificaciones por Carrera y el reporte de Becas	445

Ilustración 508 Evaluación del estado de la unión	445
Ilustración 509 Evaluación específica de la unión	445
Ilustración 510 Unión del segundo periodo del reporte de asignaturas perdidas con la unión inicial	446
Ilustración 511 Evaluación del estado de la unión completa	446
Ilustración 512 Evaluación específica de la unión completa	446
Ilustración 513 Unión del tercero periodo reporte de Calificaciones por Carrera y el reporte de Becas	447
Ilustración 514 Evaluación del estado de la unión	447
Ilustración 515 Evaluación específica de la unión	447
Ilustración 516 Análisis de la variable etnia	448
Ilustración 517 Unión del tercero periodo del reporte de asignaturas perdidas con la unión inicial	448
Ilustración 518 Evaluación del estado de la unión completa	449
Ilustración 519 Evaluación específica de la unión completa	449
Ilustración 520 Unión del cuarto periodo reporte de Calificaciones por Carrera y el reporte de Becas	450
Ilustración 521 Evaluación del estado de la unión	450
Ilustración 522 Evaluación específica de la unión	450
Ilustración 523 Análisis de la variable etnia	451
Ilustración 524 Unión del cuarto periodo del reporte de asignaturas perdidas con la unión inicial	451
Ilustración 525 Evaluación del estado de la unión completa	451

Ilustración 526	Evaluación específica de la unión completa	452
Ilustración 527	Vista del primero periodo unificado.....	452
Ilustración 528	Vista del segundo periodo unificado	452
Ilustración 529	Vista del tercer periodo unificado	453
Ilustración 530	Vista del cuarto periodo unificado	453
Ilustración 531	Código que permite la identificación de variables	454
Ilustración 532	Resultados del primer y segundo periodos unificados	455
Ilustración 533	Resultados del tercer y cuarto periodos unificados	456
Ilustración 534	Comparativa del primer periodo unificado	457
Ilustración 535	Comparativa del segundo periodo unificado.....	457
Ilustración 536	Comparativa del tercer periodo unificado	458
Ilustración 537	Comparativa del primer periodo unificado	458
Ilustración 538	Concatenación de todos los periodos	460
Ilustración 539	Evaluación visual de la unión.....	460
Ilustración 540	Evaluación avanzada de la unión	460
Ilustración 541	Vista previa del conjunto de datos unificado	461
Ilustración 542	Eliminación de duplicados del conjunto de datos unificado	462
Ilustración 543	Cantidad de registros en el conjunto de datos	463
Ilustración 544	Estadísticas descriptivas de los datos	463
Ilustración 545	Análisis de cantidad de valores faltantes.....	464
Ilustración 546	Análisis del porcentaje de los valores faltantes.....	464
Ilustración 547	Pare 1 de los resultados del análisis del porcentaje de los valores faltantes	465

Ilustración 548 Parte final de los resultados del análisis del porcentaje de los valores faltantes	466
Ilustración 549 Análisis de las variables periodo, facultad, area_asignatura y carrera.	469
Ilustración 550 Creación de un nuevo conjunto de datos y eliminación de valores faltantes de la variable etnia.....	472
Ilustración 551 Eliminación de valores faltantes de las variables periodo, facultad, area_asignatura y carrera	473
Ilustración 552 Eliminación de valores faltantes de la variable docente	473
Ilustración 553 Verificación del proceso de eliminación de valores faltantes	473
Ilustración 554 Parte 1 de la imputación de valores faltantes en la variable nivel.....	474
Ilustración 555 Parte 2 de la imputación de valores faltantes en la variable nivel.....	475
Ilustración 556 Parte 1 de la imputación de valores faltantes en la variable aprobado.	476
Ilustración 557 Parte final de la imputación de valores faltantes en la variable aprobado	476
Ilustración 558 Parte 1 de la imputación de valores faltantes en la variable credits...	477
Ilustración 559 Parte final de la imputación de valores faltantes en la variable credits	477
Ilustración 560 Verificación de la imputación de valores faltantes	478
Ilustración 561 Eliminación de los valores faltantes de la variable credits.....	478
Ilustración 562 Parte 1 de la revisión final del conjunto de datos.....	479
Ilustración 563 Parte final de la revisión final del conjunto de datos	479
Ilustración 564 Verificación y corrección de las variables estudiante y periodo	480
Ilustración 565 Verificación y corrección de la variable etnia.....	480

Ilustración 566	Verificación y corrección de la variable nivel	481
Ilustración 567	Análisis avanzado de la variable nivel	481
Ilustración 568	Eliminación de los registros con 'EL' y 'ELECT'	482
Ilustración 569	Vista previa del cambio.....	482
Ilustración 570	Verificación de las variables de segundas_matriculas y terceras_matriculas	483
Ilustración 571	Verificación de la variable facultad	484
Ilustración 572	Verificación de la variable docente.....	484
Ilustración 573	Verificación de la variable area_asignatura	485
Ilustración 574	Cambio de tipo de dato de la variable area_asignatura y su verificación	485
Ilustración 575	Verificación de la variable carrera	486
Ilustración 576	Parte 1 del resultado de la verificación de la variable carrera.....	487
Ilustración 577	Parte final del resultado de la verificación de la variable carrera	488
Ilustración 578	Verificación de la variable aprobado	489
Ilustración 579	Verificación de la variable creditos.....	489
Ilustración 580	Vista del conjunto de datos	490
Ilustración 581	Vista de los tipos de datos del conjunto de datos.....	490
Ilustración 582	Estandarización con StandardScaler	491
Ilustración 583	Evaluación de la estandarización	491
Ilustración 584	Normalización con MinMaxScaler	492
Ilustración 585	Evaluación de la normalización	492
Ilustración 586	Codificación y evaluación de la variable aprobado	493

Ilustración 587	Aplicación de One-Hot Encoding	493
Ilustración 588	Aplicación de Target Encoding para la variable docente.....	493
Ilustración 589	Revisión de las dimensiones	494
Ilustración 590	División de los datos en entrenamiento y prueba.....	496
Ilustración 591	Modelo de regresión logística	497
Ilustración 592	Random Forest	499
Ilustración 593	Support Vector Machines (SVM)	500
Ilustración 594	Parte 1 de la primera propuesta de Red Neuronal.....	501
Ilustración 595	Parte 2 de la primera propuesta de Red Neuronal.....	501
Ilustración 596	Parte final de la primera propuesta de Red Neuronal.....	502
Ilustración 597	Parte 1 de la segunda propuesta de red neuronal	503
Ilustración 598	Parte 2 de la segunda propuesta de red neuronal	503
Ilustración 599	Parte final de la segunda propuesta de red neuronal	504
Ilustración 600	Parte 1 de la tercera propuesta de red neuronal.....	505
Ilustración 601	Parte 2 de la tercera propuesta de red neuronal.....	505
Ilustración 602	Parte final de la tercera propuesta de red neuronal	506
Ilustración 603	Parte 1 de la cuarta propuesta de red neuronal.....	507
Ilustración 604	Parte 2 de la cuarta propuesta de red neuronal.....	507
Ilustración 605	Parte final de la cuarta propuesta de red neuronal	508
Ilustración 606	Parte 1 de la quinta propuesta de red neuronal.....	509
Ilustración 607	Parte 2 de la quinta propuesta de red neuronal.....	510
Ilustración 608	Parte 3 de la quinta propuesta de red neuronal.....	510
Ilustración 609	Parte final de la quinta propuesta de red neuronal	511

Ilustración 610 Parte 1 de la sexta propuesta de red neuronal	512
Ilustración 611 Parte 2 de la sexta propuesta de red neuronal	513
Ilustración 612 Parte final de la sexta propuesta de red neuronal	514
Ilustración 613 Parte 1 de la séptima propuesta de red neuronal	515
Ilustración 614 Parte 2 de la séptima propuesta de red neuronal	516
Ilustración 615 Parte final de la séptima propuesta de red neuronal	517
Ilustración 616 Creación de la quinta propuesta de red neuronal	531
Ilustración 617 Primera parte del entrenamiento de la quinta propuesta de red neuronal	531
Ilustración 618 Parte final del entrenamiento de la quinta propuesta de red neuronal .	532
Ilustración 619 Evaluación de la quinta propuesta de red neuronal	533
Ilustración 620 Exportación del modelo de la quinta propuesta de red neuronal	534
Ilustración 621 Creación de una copia de los datos antes de la estandarización y normalización.....	534
Ilustración 622 Estandarización de variables numéricas.....	535
Ilustración 623 Normalización de variables numéricas	535
Ilustración 624 Vista previa de los datos estandarizados y normalizados	535
En contraste, la copia del conjunto de datos antes de ser procesado tiene el siguiente contenido: Ilustración 625 Vista previa de los datos sin estandarizar ni normalizar	536
Ilustración 626 Parte 1 de la visualización del contraparte de los datos estandarizados y normalizados de la variable credits	537
Ilustración 627 Parte final de la visualización del contraparte de los datos estandarizados y normalizados de la variable credits.....	538

Ilustración 628 Visualización del contraparte de los datos estandarizados y normalizados de la variable segundas_matriculas	539
Ilustración 629 Visualización del contraparte de los datos estandarizados y normalizados de la variable terceras_matriculas	539
Ilustración 630 Creación de diccionarios con los valores únicos de las variables numéricas y sus valores estandarizados y normalizados	540
Ilustración 631 Codificación de la variable aprobado	541
Ilustración 632 Creación de una lista con las categorías de las variables categóricas con excepción de la variable docente	542
Ilustración 633 Ejecución de la técnica de One-Hot-Encoding	542
Ilustración 634 Visualización del contraparte de los datos codificados de la variable docente	544
Ilustración 635 Agregación de los valores únicos de la variable docente y sus valores codificados al diccionario “diccionario_para_exportar”	545
Ilustración 636 Código para la creación de un documento ".txt" y el guardado de todas las categorías	545
Ilustración 637 Vista previa del documento creado	546
Ilustración 638 Vista de la ubicación de los archivos donde se realizará el aplicativo web	547
Ilustración 639 Descripción del modulo e importación de "pandas"	547
Ilustración 640 Definición de la clase, el constructor y la función importar_datos	548
Ilustración 641 Definición de la función clasificar_datos	549

Ilustración 642 Definición de las funciones get_sincodificar, get_codificadas y get_datos	550
Ilustración 643 Definición de la función get_seleccionado	551
Ilustración 644 Primera parte de la definición de la función crear_propuesta_prediccion	552
Ilustración 645 Vista el formato de nombres de la variable etnia	553
Ilustración 646 Parte final de la definición de la función crear_propuesta_prediccion	553
Ilustración 647 Importación de modulos, definiciones iniciales y declaración del modelo de la página principal	554
Ilustración 648 Parte 1 de la definición del modelo de la página "predicción"	555
Ilustración 649 Parte final de la definición del modelo de la página "predicción"	556
Ilustración 650 Definición del modelo de la página "resultado"	557
Ilustración 651 Contenido del aplicativo web	558
Ilustración 652 Logo de SPRAST	559
Ilustración 653 Primera parte de la página de inciio de SPRAST	560
Ilustración 654 Parte final de la página de inciio de SPRAST	560
Ilustración 655 Parte 1 de la página de ingreso de datos del estudiante para su predicción	561
Ilustración 656 Parte 2 de la página de ingreso de datos del estudiante para su predicción	561
Ilustración 657 Parte final de la página de ingreso de datos del estudiante para su predicción	561
Ilustración 658 Página donde se puede apreciar el resultado de la predicción	562

ÍNDICE DE TABLAS

Tabla 1 Identificación de Variables del Reporte de Asignaturas Perdidas	44
Tabla 2 Diccionario de Códigos MATE (Áreas de Asignaturas)	46
Tabla 3 Diccionario de Códigos de NF (Nivel Académico)	46
Tabla 4 Identificación de Variables del Reporte de Calificaciones por Carrera.....	47
Tabla 5 Identificación de variables del reporte de Becas.....	50
Tabla 6 Identificación de PII's del Reporte de Asignaturas Perdidas	54
Tabla 7 Identificación de PII's del Reporte de Calificaciones por Carrera.....	55
Tabla 8 Identificación de PII's del Reporte de Becas	56
Tabla 9 Clasificación de Variables del Reporte de Asignaturas Perdidas	72
Tabla 10 Clasificación de Variables del Reporte de Calificaciones por Carrera.....	74
Tabla 11 Clasificación de Variables del Reporte de Becas	77
Tabla 12 Variables con un único valor y sin relevancia	122
Tabla 13 Variables con miles de valores y sin utilidad para el análisis.....	124
Tabla 14 Variables redundantes	124
Tabla 15 Variables que a pesar de no tener un análisis profundo no son relevantes	126
Tabla 16 Variables que, dado el enfoque de la investigación, son irrelevantes.....	126
Tabla 17 Variables seleccionadas como irrelevantes.....	127
Tabla 18 Determinación de irrelevancia y relevancia de variables del reporte de Asignaturas Perdidas.....	328
Tabla 19 Determinación de irrelevancia y relevancia de variables del reporte de Calificaciones por Carrera	344

Tabla 20 Determinación de Irrelevancia y Relevancia de Variables del reporte de Becas	389
Tabla 21 Selección de variables del reporte de Asignaturas Perdidas	391
Tabla 22 Selección de variables del reporte de Calificaciones por Carrera.....	391
Tabla 23 Selección de variables del reporte de Becas	392
Tabla 24 Revisión del Reporte de Asignaturas Perdidas	408
Tabla 25 Diccionario de códigos de áreas de asignaturas.....	409
Tabla 26 Revisión del Reporte de Calificaciones por Carrera.....	410
Tabla 27 Revisión del Reporte de Becas	410
Tabla 28 Resultados de la Eliminación de Duplicados del Reporte de Asignaturas Perdidas	414
Tabla 29 Resultados de la Eliminación de Duplicados del Reporte de Calificaciones por Carrera.....	415
Tabla 30 Resultados de la Eliminación de Duplicados del Reporte de Becas	415
Tabla 31 Consideraciones para la armonización de las variables del reporte de Asignaturas Perdidas.....	419
Tabla 32 Consideraciones para la armonización de las variables del reporte de Calificaciones por Carrera	420
Tabla 33 Consideraciones para la armonización de las variables del reporte de Becas	420
Tabla 34 Identificación de variables que se atribuyen a los mismos datos del reporte de Asignaturas Perdidas.....	423
Tabla 35 Identificación de variables que se atribuyen a los mismos datos del reporte de Calificaciones por Carrera	424

Tabla 36 Identificación de variables que se atribuyen a los mismos datos del reporte de Becas	424
Tabla 37 Identificación de las variables que requieren armonización	429
Tabla 38 Detalle de variables a partir de la comparativa	459
Tabla 39 Resultados de la eliminación de duplicados	462
Tabla 40 Identificación de valores faltantes	466
Tabla 41 Árbol de decisión	498
Tabla 42 Tabla de evaluaciones de los modelos	518
Tabla 43 Descripción de Variables Resultantes.....	521
Tabla 44 Resultados de los modelos realizados.....	523

CAPÍTULO I: INTRODUCCIÓN

1. Introducción

1.1. Justificación

“Cuando se habla de éxito o fracaso en la etapa universitaria, el término más comúnmente usado para medirlo es el rendimiento académico” (Dávila Cárdenes, García Artilles, Pérez Sánchez, & Gómez Déniz, 2015, pág. 29). Por lo que, al usarse el rendimiento académico para determinar el éxito o fracaso de un estudiante en términos de su vida académica resulta imperativo el desarrollo de procesos efectivos y eficientes que permitan prever estos resultados, debido a que, “Tomando cifras del Banco Mundial, a nivel de Latinoamérica y el Caribe, el abandono universitario en personas de 25 a 29 años, durante la última década, ha sido de alrededor del 21%, ascendiendo al 30 % para el caso de Ecuador” (Ferreryra et al, 2017 como se citó en Zumárraga Espinosa & Cevallos Pozo, 2022, pág. 3). Esto, a razón de “el rendimiento académico como un determinante fundamental del fracaso y abandono estudiantil de programas de estudio de tercer nivel.” (Zumárraga Espinosa & Cevallos Pozo, 2022, pág. 3).

Además, “El rendimiento académico en estudiantes universitarios representa un indicador estratégico para la valoración de la calidad educativa en la educación superior, y es el resultado de la unión de diferentes factores que intervienen en la vida académica del estudiante” (como se citó en Vargas, 2013). Es decir, el rendimiento académico es una métrica importante para evaluar el éxito de los estudiantes universitarios en su educación, “es por esta razón que se han utilizado técnicas de aprendizaje automático para la creación de modelos predictivos empleados en la predicción del rendimiento académico con el fin de identificar los factores que influyen en el proceso de enseñanza” (Rico Páez & Gaytán Ramírez, 2022), por lo que, el diseño de un modelo de predicción de esta métrica puede proporcionar una herramienta valiosa para ayudar a los

estudiantes, profesores y administradores universitarios a abarcar de mejor manera los factores que afectan el rendimiento académico.

Adicionalmente, el desarrollo de un modelo de machine learning para predecir el rendimiento académico de los estudiantes universitarios resulta una estrategia innovadora para abordar este problema, ya que “las herramientas estadísticas clásicas pueden no funcionar adecuadamente con grandes cantidades de datos y con varias características de estudiantes” (Romero y Ventura, 2012 como se citó en Rico Páez & Gaytán Ramírez, 2022). Además, esta estrategia es reciente en países de Latinoamérica (Rico Páez & Gaytán Ramírez, 2022). Es decir, con un modelo de predicción del rendimiento académico, “La reprobación (...) puede ser disminuida si se tiene información oportuna de los estudiantes (...) y de su desempeño académico.” (Rico Páez & Gaytán Ramírez, 2022)

Por otra parte, dado que “En México existe poco desarrollo en la construcción de este tipo de modelos con técnicas de aprendizaje automático debido, principalmente, al desconocimiento de la metodología para llevarlo a cabo.” (Rico Páez & Gaytán Ramírez, 2022). El uso de una metodología como CRISP-DM es fundamental para garantizar una alta probabilidad de éxito en el desarrollo del modelo, ya que, “es un método probado para orientar sus trabajos de minería de datos” (IBM, 2021). Que aplica igualmente para trabajos de aprendizaje automático por la razón de que proporciona un marco estructurado y sistemático para la identificación de variables relevantes, su limpieza, comprensión, preparación, modelado y evaluación.

En conclusión, el diseño de un modelo de predicción del rendimiento académico universitario mediante técnicas de machine learning puede proporcionar una herramienta valiosa para mejorar la educación universitaria, identificar áreas de mejora y ofrecer apoyo y recursos a

los estudiantes que lo necesitan; además, el uso una metodología como CRISP-DM para llevar a cabo el modelo permitirá una alta probabilidad de éxito en cuestión de predecir si un estudiante aprobará o reprobará una asignatura que está considerando cursar o está por cursar.

1.1.1. Alcance

Este proyecto se centrará en el desarrollo de un modelo de machine learning para predecir el rendimiento académico de los estudiantes universitarios en términos de aprobación o reprobación de materias utilizando la metodología CRISP-DM. Se recopilarán datos históricos de desempeño académico de todos los estudiantes universitarios y se utilizarán técnicas de minería de datos para identificar las variables más relevantes para la predicción, además, se seleccionará y entrenará un modelo de machine learning y se evaluará su desempeño utilizando métricas de evaluación pertinentes. Finalmente, se desplegará el modelo para su uso en el mundo real de manera simple, es decir, este proyecto no abordará la implementación del modelo de machine learning en un sistema de gestión educativa o la integración del modelo en el proceso de toma de decisiones de los profesores.

1.2. Planteamiento del problema

La mejora del rendimiento académico de los estudiantes universitarios es un objetivo clave en la educación superior. Sin embargo, hasta el momento, no se ha identificado un modelo de machine learning que pueda predecir con precisión el rendimiento académico futuro de un estudiante a partir de registros académicos pasados. Este problema se agrava debido a la falta de datos precisos y completos sobre el desempeño académico previo de los estudiantes, lo que dificulta la identificación de las variables relevantes, relaciones y tendencias en los datos de rendimiento académico. Además, la falta de preparación adecuada de los datos de desempeño

académico y la selección de un modelo de machine learning adecuado son obstáculos importantes para la creación de un modelo de predicción del rendimiento académico.

De lo discutido se puede identificar el siguiente problema principal:

No se ha identificado un modelo de machine learning que, a partir de registros académicos pasados, sea capaz de predecir el rendimiento académico futuro de un estudiante con un alto grado de precisión.

Y los siguientes problemas secundarios:

- No se han recolectado de manera precisa y completa los datos de desempeño académico previo de los estudiantes
- No se ha determinado un conjunto de datos de rendimiento académico en el que se identifiquen variables relevantes, relaciones y tendencias de los datos del rendimiento académico.
- No se han preparado los datos de desempeño académico previo de los estudiantes para ser utilizados en modelos de aprendizaje automático.
- No se ha seleccionado un modelo de machine learning adecuado para predecir el rendimiento académico.
- No se sabe qué tan preciso es el modelo de machine learning seleccionado.

1.3. Objetivos

1.3.1. Objetivo General

Desarrollar y validar un modelo de machine learning utilizando la metodología CRISP-DM que sea capaz de predecir si un estudiante universitario aprobará o reprobará una materia, en función de variables como el desempeño académico previo

1.3.2. *Objetivos Específicos*

Con base en el planteamiento del problema, se pueden definir los siguientes objetivos específicos:

- Identificar un conjunto de datos para el modelado del rendimiento académico.
- Explorar el conjunto de datos para el modelado del rendimiento académico.
- Preparar los datos de desempeño académico.
- Desarrollar modelos de machine learning que sean capaces de predecir el rendimiento académico.
- Evaluar los modelos de machine learning
- Desplegar el modelo de machine learning mejor evaluado

1.4. Estructura del trabajo

El trabajo de titulación presente ha sido realizado utilizando la metodología CRISP-DM que será explicada en el capítulo 2, sin embargo, de manera general la metodología consiste en 6 fases que son las siguientes:

1. Fase de comprensión del negocio
2. Fase de comprensión de los datos
3. Fase de preparación de los datos
4. Fase de modelado
5. Fase de evaluación
6. Fase de despliegue

Cada fase tiene un alto nivel de importancia y afectará de igual manera al desarrollo del modelo de predicción del rendimiento académico planteado, dicho esto, la estructura de este trabajo de

titulación está estrechamente ligado con las distintas fases existentes en la metodología. No obstante, dado que la fase de comprensión del negocio tiene como cometido “alinearse los objetivos del proyecto de data mining con los objetivos del negocio. Tratando así de evitar embarcarnos en un proyecto de minería de datos que no produzca ningún efecto real en la organización.” (Rueda, 2019). Es evidente la necesidad de exponer la fase de comprensión del negocio en el primer capítulo de este trabajo de titulación con el objetivo de definir los objetivos del modelo de predicción del rendimiento académico más allá de los determinados, determinar el criterio de éxito del modelo y la investigación, determinar las partes interesadas, y, finalmente, determinar posibles riesgos y desafíos.

A partir de lo mencionado, la comprensión del negocio es la primera fase de la metodología CRISP-DM, la cual será aplicada en el desarrollo de un modelo de machine learning para predecir el rendimiento académico de los estudiantes universitarios en términos de aprobación o reprobación de materias, como parte de este trabajo de titulación. En esta fase, se tiene como objetivo definir los objetivos específicos del modelo y la investigación más allá de los especificados anteriormente, y comprender completamente el problema que se busca resolver (Chapman, y otros, 1999, pág. 13). Lo que implica determinar los criterios de éxito y el resultado deseado, así como identificar a las partes interesadas y sus intereses; de esta manera, se puede garantizar que el modelo de predicción esté alineado con los objetivos del proyecto y pueda brindar información relevante que satisfaga las necesidades de las partes interesadas.

1.4.1. Definición de objetivos de la investigación y entendimiento del problema

El objetivo principal de esta investigación es desarrollar un modelo predictivo para el rendimiento de los estudiantes universitarios a través del uso de técnicas de aprendizaje automático, este modelo ayudará a determinar si un estudiante es probable que apruebe o

repruebe una asignatura que el estudiante está considerando cursar o está por cursar, esto, a partir de datos históricos de todos los estudiantes que han cursado aquellas determinadas asignaturas; por otra parte, al desarrollar un modelo de esta manera, las universidades podrían identificar a los estudiantes que corren el riesgo de suspender desde temprano y proporcionarles el apoyo y los recursos necesarios para mejorar su rendimiento, lo que implica que no solo se beneficiará a los estudiantes, sino que también ayudará a las universidades a mejorar sus tasas de retención y mantener altos estándares académicos.

Adicionalmente, además de desarrollar el modelo predictivo, la investigación tiene como objetivo investigar los factores que afectan el rendimiento académico de los estudiantes y su importancia relativa en el modelo, este análisis proporcionará información sobre las causas subyacentes del éxito o fracaso de los estudiantes y ayudará a identificar áreas en las que las universidades pueden mejorar su apoyo a los estudiantes.

Finalmente, la investigación tiene como objetivo contribuir al cuerpo existente de conocimiento en el campo del machine learning en entornos educativos y de análisis predictivo, es decir, el modelo y las ideas obtenidas de esta investigación se sumarán a la comprensión de cómo se pueden aplicar las analíticas de datos en la educación para mejorar el rendimiento y el apoyo del estudiante; por añadidura, al compartir estos hallazgos con otros investigadores y educadores, la investigación ayudará a avanzar en el campo y promover prácticas basadas en evidencia en la educación, por lo que, se puede concluir que el objetivo es ayudar a las universidades a brindar a sus estudiantes la mejor educación y apoyo posible, lo que conduce a mejores resultados para todos.

Es decir, se definieron los siguientes objetivos específicos de la investigación:

- 1) El objetivo principal es desarrollar un modelo predictivo para el rendimiento de los estudiantes universitarios con el fin de determinar si un estudiante aprobará o reprobará una materia haciendo uso de datos históricos de todos los estudiantes que han cursado aquella determinada asignatura antes del considerado.
- 2) Realizar una investigación y análisis de los factores que afectan el rendimiento estudiantil y su importancia relativa en el modelo, con el objetivo de brindar nuevos enfoques para investigaciones futuras o aplicaciones prácticas.
- 3) Contribuir al cuerpo existente de conocimiento en el campo del machine learning en entornos educativos y de análisis predictivo.

1.4.2. Determinación del criterio de éxito y el resultado deseado

Los criterios de éxito para el modelo predictivo incluyen la demostración de un alto nivel de precisión y rendimiento en la tarea de predecir el rendimiento académico de los estudiantes considerando que se utilizarán datos reales.

Por otra parte, el modelo deberá ser construido utilizando metodologías rigurosas y estar bien documentado, debido a que esto asegurará que otros investigadores puedan comprender, reproducir y construir sobre el trabajo realizado; adicionalmente, la documentación adecuada también permitirá que el modelo se actualice y mejore con el tiempo a medida que se disponga de nuevos datos.

Por otra parte, los hallazgos de la investigación deben contribuir con nuevos conocimientos o métodos en el campo del machine learning para la predicción del rendimiento académico y ser de potencial interés para futuros investigadores o profesionales, en otras palabras, estos conocimientos deben proporcionar una base para investigaciones adicionales en el

área, refinar el modelo predictivo, identificar nuevos factores que afectan el rendimiento estudiantil y desarrollar nuevas intervenciones para mejorar los resultados de los estudiantes.

Es decir, se determinaron los siguientes criterios de éxito:

- 1) El modelo predictivo debe demostrar un alto nivel de precisión y rendimiento, considerando el uso de datos reales.
- 2) El modelo debe construirse utilizando metodologías rigurosas y estar bien documentado, asegurando que, otros investigadores puedan comprender, reproducir y construir sobre su trabajo. En esta ocasión, se utilizará la metodología CRISP-DM.
- 3) Los hallazgos de la investigación deben contribuir con nuevos conocimientos o métodos al campo y ser de potencial interés para futuros investigadores o profesionales.

1.4.3. Identificación de las partes interesadas

En primera instancia, la primera parte interesada serán los investigadores que deseen nutrir sus conocimientos con nuevos enfoques y desarrollos dentro del área del entorno educativo, en específico, con respecto al uso de machine learning para predecir el rendimiento académico, aunque, esto también implica para los resultados de los análisis realizados sobre el rendimiento académico de los estudiantes de la Pontificia Universidad Católica del Ecuador.

Por otra parte, la Pontificia Universidad Católica podría ser otra parte interesada en los éxitos de esta investigación, debido a que, dado que los análisis y modelo de predicción fueron realizados a partir de datos reales de los estudiantes de la mencionada universidad, esta podría beneficiarse inmensamente en términos de generar nuevos procesos para tratar problemas relacionados con el bajo rendimiento académico.

En conclusión, se destacaron dos principales partes interesadas:

- Futuros investigadores del área del rendimiento académico, machine learning y modelos predictivos.
- La Pontificia Universidad Católica del Ecuador.

1.4.4. Identificación de posibles riesgos y desafíos

El primer riesgo que puede presentarse es con respecto a la calidad de los datos obtenidos, puesto que, a pesar de que se realice un exhaustivo análisis sobre estos datos y se destaquen variedad de conclusiones que aparentan certitud, si los datos presentan sesgos, poca fidelidad o inconsistencias, podría afectar directamente a los resultados del modelo de predicción del rendimiento académico, ya que, los resultados de los análisis realizados no tendrían una real validez.

Por otra parte, otro riesgo que podría presentarse es que sin importar la hipótesis resultante de la fase de comprensión de los datos con los que se formulará el modelo de predicción de rendimiento académico, no sea factible el cumplir con los criterios de éxito del modelo con respecto a la precisión.

De manera general, se han destacado los siguientes dos riesgos:

- Riesgos relacionados a la calidad, fidelidad y consistencia de los datos.
- Riesgos relacionados a que los datos obtenidos, no sean útiles con respecto al rendimiento académico.

1.5. Notas respecto a los códigos realizados en el trabajo

A lo largo de la lectura de este documento se podrá visualizar una gran cantidad de funciones, métodos y códigos que permitirán el cumplimiento de los objetivos del presente trabajo de titulación, sin embargo, es necesario mencionar que todos los códigos que fueron realizados durante esta investigación son de autoría del investigador, por supuesto, con excepción de los módulos y librerías utilizadas, y debidamente mencionadas en sus importaciones, en cuyo aspecto, los derechos van a sus respectivos autores.

CAPÍTULO II: REVISIÓN SISTEMÁTICA DE LITERATURA

2. Revisión sistemática de literatura

2.1. Rendimiento académico en estudiantes universitarios

Como punto de partida, el rendimiento académico es “el informe que logra el estudiante, definido por una nota global o calificación escolar, la cual es considerada como el reflejo del aprendizaje y de los conocimientos adquiridos” (Lafourcade, 1962 como se citó en Huerta Reyes, Pérez García, López González, Uscanga Hermida, & Ríos Carrillo, 2023, pág. 6). Sin embargo, el concepto también aplica para el contexto universitario teniendo en cuenta que, durante cada periodo académico, los estudiantes culminan con una calificación final que refleja su aprendizaje, de hecho, de manera general, “El rendimiento académico tiene un carácter complejo y multidimensional que se relaciona directamente con los resultados de aprendizaje, sean estos autónomos o colaborativos, que por lo general se reflejan en las calificaciones obtenidas por el estudiantado.” (García, 2015 como se citó en (Borja Naranjo, Martínez Benítez, Barreno Freire, & Garo Jácome, 2021).

A partir de este concepto, “el rendimiento académico en universitarios, en su mayoría se ve afectado por el estrés, puesto que todas las actividades implican la inversión del tiempo, esfuerzo, incluso hasta el recurso económico” (como se citó en Huerta Reyes, Pérez García, López González, Uscanga Hermida, & Ríos Carrillo, 2023, pág. 6). Lo que indica que hay factores que influyen directa e indirectamente en el rendimiento académico de los estudiantes universitarios, de hecho, “los estudiantes de medicina enfrentan situaciones que afectan su salud mental y rendimiento académico, por lo que en un estudiante susceptible puede manifestar ansiedad, depresión y estrés” (González-Olaya, 2014, como se citó en Huerta Reyes, Pérez García, López González, Uscanga Hermida, & Ríos Carrillo, 2023, pág. 6). Esto quiere decir,

que además de que existen factores que influyen en el rendimiento académico de los estudiantes universitarios, estos también repercuten negativamente en la salud de estos.

Por consiguiente, existen factores internos y externos que afectan en determinada manera a los estudiantes universitarios, como, por ejemplo, condiciones relacionadas a su familia, estado financiero, nivel de educación máximo alcanzado por sus padres, la complejidad inherente del estudio, la motivación, la depresión, el colegio de provino, etc. (Martínez Benítez, Borja Naranjo, Haro Jácome, & Barreno Freire, 2020, pág. 4). Además, también podría influir que los estudiantes se encuentran en una época de cambios drásticos y significativos como el paso de la adolescencia a la adultez y posterior actividad laboral (Borja Naranjo, Martínez Benítez, Barreno Freire, & Garo Jácome, 2021, pág. 2).

Adicionalmente, “de acuerdo con los estudios realizados, se ha evidenciado que un buen desempeño académico no puede ser garantizado solo por la capacidad intelectual de los individuos pues la acción de otras variables no cognitivas puede determinar un desempeño diferente” (Ruíz Dodobara, 2005, como se citó en Zumárraga Espinosa & Cevallos Pozo, 2022, pág. 3). Lo que concreta que no existe un factor único que puede determinar el rendimiento académico de un estudiante universitario dado.

En conclusión, la labor de determinar el rendimiento académico en estudiantes universitarios es una tarea de extrema importancia para las instituciones educativas, ya que sus estudiantes pueden perjudicarse por una mala gestión de los procesos que se llevan a cabo en las respectivas universidades, dichos perjuicios pueden ser tanto en términos académicos como en su salud mental y física; por lo que, además de repercutir en la reputación de los estudiantes para el resto de su vida laboral y académica, también afecta directamente a la imagen pública que poseen las universidades en términos de excelencia académica.

2.2. Modelos de Machine Learning para la predicción del rendimiento académico

“El progreso humano está basado en la invención de procedimientos y mecanismos para la resolución de problemas concretos del quehacer cotidiano. Desde las primeras técnicas para encender fuego, hasta las complejas máquinas del mundo moderno” (Ordóñez, 2007, pág. 1). Dicho de otra manera, el ingenio humano ha sido capaz de progresivamente recurrir a nuevas técnicas y métodos que le permiten agendar los mismos problemas de mejor manera, pues, es evidente como en las últimas décadas se ha evidenciado un crecimiento tecnológico extraordinario comúnmente conocido como “tsunami tecnológico”. (Hernández, 2022). Esta rápida evolución ha sido fácilmente visualizada sobre todo en el área de la inteligencia artificial.

A partir de esta premisa, asimismo se ha presenciado un crecimiento del uso de la tecnología dentro del ámbito académico donde es visible como progresivamente se ha aumentado su uso para fomentar prácticas adecuadas en la enseñanza y aprendizaje (como se citó en Rico Páez & Gaytán Ramírez, 2022, pág. 2). Por lo que se infiere su crecimiento en uso dentro del análisis y predicción del rendimiento académico, ya que con el avance de las ramas de la inteligencia artificial se ha permitido la oportunidad de agendar el problema del rendimiento académico con enfoques innovadores.

Debido a que, el rendimiento académico, como se ha ido explicando continuamente, es un término que su sentido suele variar de acuerdo con varios factores, es decir, el rendimiento académico de un determinado estudiante es muy probable que no corresponda a otro estudiante a pesar de que se encuentre en la misma universidad, colegio, país, etc. (como se citó en Rico Páez & Gaytán Ramírez, 2022, pág. 2). Dicho de otra manera, el rendimiento académico resulta un área complicada para realizar un análisis estadístico tradicional para determinar si un estudiante tiene mayor o menor probabilidad de desempeñarse aceptablemente en una determinada

asignatura, puesto que, como se mencionó en la justificación, los métodos y procesos estadísticos tradicionales decrecen su utilidad a medida que aumentan la cantidad de datos y factores que influyen sobre estos datos (Romero y Ventura, 2012 como se citó en Rico Páez & Gaytán Ramírez, 2022).

Es por lo que, dado el avance tecnológico y el ingenio humano, el aprendizaje automático o machine learning “se ha comenzado a emplear en el diseño de modelos predictivos del rendimiento académico a partir de factores o características de estudiantes” (Romero y Ventura, 2010 como se citó en Rico Páez & Gaytán Ramírez, 2022, pág. 3). Teniendo en cuenta que esta área de la inteligencia artificial permite a un sistema aprender por su cuenta a partir de datos en lugar de ser explícitamente enseñado por un programador (IBM, s.f.). Es notoria la razón por la que el aprendizaje automático ha comenzado a ser plenamente utilizado en el campo del estudio y entendimiento del rendimiento académico, pues, el desarrollo de modelos de predicción del rendimiento académico permite identificar aquellos factores que influyen directamente en los estudiantes. (Rico Páez & Gaytán Ramírez, 2022)

2.3. Metodología CRISP-DM

Para empezar, CRISP-DM es un acrónimo de Cross Industry Standard Process for Data Mining (Hotz, 2023). No obstante, a pesar de que no existe una traducción oficial en español para la terminología, un acercamiento prudente podría ser: “Proceso Estándar de Minería de Datos para las Industrias”; de todas maneras, CRISP-DM es un modelo de proceso que sirve como base para procesos de ciencia de datos, el mismo, publicado en 1999 con el fin de estandarizar los procesos de minería de datos, se ha convertido en la metodología mayormente utilizada para proyectos de ciencia de datos. (Hotz, 2023). Es decir, la metodología CRISP-DM es un proceso estándar para desarrollar proyectos de ciencia de datos sin importar el tipo de

industria, en esta ocasión se presenta un proyecto de ciencia de datos enfocado en el desarrollo de un modelo de predicción, y, la industria consiste en la educación universitaria.

De manera general, la metodología CRISP-DM consiste en 6 fases:

1. Fase de comprensión del negocio
2. Fase de comprensión de los datos
3. Fase de preparación de los datos
4. Fase de modelado
5. Fase de evaluación
6. Fase de despliegue

2.3.1. Entendimiento del negocio

Para empezar, como se ha explicado brevemente en el capítulo 1, la primera fase se concentra en el entendimiento de los objetivos y requerimientos específicos desde una perspectiva del negocio (Chapman, y otros, 1999, pág. 13). Es decir, esta fase se dedica principalmente en comprender que es lo que se debe realizar y esperar del proceso de creación del modelo de predicción de rendimiento académico, por lo que, esta fase generalmente es consistida por las siguientes actividades:

- En primer lugar, en concordancia con el presente trabajo, se debe entender a detalle las necesidades del proyecto desde una perspectiva de la investigación y del desarrollo del modelo (Hotz, 2023). En otras palabras, se deben identificar objetivos para la investigación y proceso de desarrollo del modelo de predicción de rendimiento académico más allá de los planteados para el trabajo de titulación.

- Se debe definir los criterios de éxito y los resultados deseados del proceso de desarrollo del modelo. (Hotz, 2023). Estos criterios deberán ser descritos en sentido del éxito y utilidad de los resultados del proyecto para el negocio. (Chapman, y otros, 1999, pág. 17). Sin embargo, también es necesario definir estos criterios no únicamente basados en términos del negocio, sino en cuestiones intrínsecas del modelo planteado. (Hotz, 2023).
- Es necesario también identificar a quienes podría interesar los resultados del proceso de desarrollo del modelo de predicción, es decir, las partes interesadas del proyecto. (CRISP-DM Guide, como se citó en Hotz, 2023).
- Finalmente, es necesario determinar los riesgos presentes e inherentes del proceso de desarrollo del modelo planteado. (Hotz, 2023). Para esto, se debe enlistar los posibles riesgos y eventualidades que podrían causar que el proyecto fracase. (Chapman, y otros, 1999, pág. 18)

2.3.2. Entendimiento de los datos

La segunda fase corresponde al entendimiento de los datos y comienza con la recolección de los datos iniciales para la investigación con los que se realizarán actividades para que el investigador se familiarice con los datos e identifique cualidades específicas de los mismos (Chapman, y otros, 1999, pág. 14). Es decir, en esta fase es necesario realizar un análisis exploratorio que permita al investigador detectar variables relevantes, irrelevantes, comportamientos sesgados, etc. Con la finalidad de formular una hipótesis, en esta ocasión, para el modelo de predicción del rendimiento académico.

Sin embargo, es necesario comentar que, si se adquieren múltiples fuentes de datos, la integración es un problema adicional que puede ser tratado en la presente fase o en la siguiente

fase de preparación de los datos. (Chapman, y otros, 1999, pág. 20). En resumen, como investigador, se tiene la potestad de decidir en qué momento realizar la integración de los datos en el caso de que se hayan obtenido varias fuentes de estos.

Adicionalmente, es necesario describir los resultados de la tarea de la exploración de los datos haciendo uso de gráficos y diagramas que permitan indicar características propias de los datos analizados. (Chapman, y otros, 1999, pág. 21). Esto, con la finalidad de presentar una hipótesis lo más aproximada a la realidad, pues, será formulada en base a las conclusiones de los análisis llevados a cabo.

2.3.3. *Preparación de los datos*

Esta tercera fase comprende a todas las actividades que serán llevadas a cabo sobre los datos obtenidos con la finalidad de construir un conjunto de datos final con todos los datos con los que será entrenado el modelo de predicción de rendimiento académico. (Chapman, y otros, 1999, pág. 14). Es decir, se aplicarán técnicas como la limpieza de datos, la construcción o formateo de datos, la integración de datos en el caso de presentar múltiples fuentes de datos y la normalización y estandarización. (Hotz, 2023).

2.3.4. *Modelado*

La cuarta fase corresponde al modelado, aquí, se llevarán a cabo varias técnicas de modelado seleccionadas que sean apropiadas para tanto los datos obtenidos a partir de la fase de preparación como para los objetivos del modelo. (Chapman, y otros, 1999, pág. 14). En esta fase es habitual la construcción de varios modelos basados en diferentes técnicas, de manera general, esta fase consiste en la selección de técnicas de modelado, la generación de un conjunto de pruebas y entrenamiento, la construcción del modelo, y la evaluación de estos. (Hotz, 2023)

2.3.5. Evaluación

La fase de evaluación continua justo después de la creación de los modelos y su objetivo es determinar el modelo más adecuado para agendar la problemática y aproximarse a los objetivos de la investigación. (Chapman, y otros, 1999, pág. 14). Es decir, para lograr la decisión descrita, se deben evaluar los resultados considerando si se cumplen los criterios de éxito definidos y la evaluación del proceso con la finalidad de determinar si es necesario realizar un nuevo proceso de modelado o desplegar el modelo más idóneo. (Hotz, 2023).

2.3.6. Despliegue

“Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise” [Según los requisitos, la fase de implementación puede ser tan simple como generar un informe o tan compleja como implementar un proceso de minería de datos repetible en toda la empresa.] (CRISP-DM Guide, como se citó en Hotz, 2023). Sin embargo, a pesar de que se mencione que la fase de despliegue tiene su dificultad en función de los requerimientos del proyecto, es necesario tener en cuenta que, sin importar la forma de despliegue, un modelo no es útil a menos de que los interesados puedan acceder a sus resultados. (Hotz, 2023). Por lo que se refleja la necesidad de desplegar acertadamente los resultados del modelo.

CAPÍTULO III: MARCO TEÓRICO

3. Marco Teórico

3.1. Aprendizaje supervisado y no supervisado

Ambos tipos de aprendizaje son ramas del machine learning, en específico, el aprendizaje supervisado es “un método de análisis de datos que utiliza algoritmos que aprenden iterativamente de los datos para permitir que los ordenadores encuentren información escondida sin tener que programar de manera explícita dónde buscar” (TIBCO, s.f.). Sin embargo, la principal diferencia con respecto al aprendizaje no supervisado es que el que el aprendizaje supervisado se enfoca en áreas cuyo resultado ya se conoce, por el contrario, el aprendizaje no supervisado es utilizado cuando no se conoce los resultados de determinados casos, por ejemplo, no se conoce cuando existen fraudes bancarios, por lo que el “el aprendizaje no supervisado es un tipo de Machine Learning que se utiliza para identificar nuevos patrones y detectar anomalías. Los datos que se introducen en los algoritmos de aprendizaje no supervisados no están etiquetados.” (TIBCO, s.f.). Por lo que, para definir qué tipo de aprendizaje se utilizará para determinados problemas de machine learning, es importante notas si existe una variable comúnmente conocida como “variable objetivo”.

3.2. Algoritmos de machine learning utilizados en el estudio

Los algoritmos y técnicas que fueron utilizados en el desarrollo de este trabajo de titulación fueron seleccionados principalmente debido al objetivo del modelo planteado, sin embargo, también se consideraron otros factores que serán hablados a detalle en el capítulo 4.3. Las técnicas utilizadas fueron:

- Regresión Logística
- Árbol de decisión

- Random Forest
- Support Vector Machines (SVM)
- Redes Neuronales

3.2.1. Regresión Logística

La regresión logística es una técnica de machine learning que utiliza funciones matemáticas con la finalidad de descubrir una relación entre dos factores, y utilizarla para predecir el valor de uno de estos. (Amazon, s. f.) Por lo general, la regresión logística cumple con la predicción de un número no infinito de valores y es ideal para modelos binarios como el predecir el rendimiento académico de un estudiante en términos de aprobación o reprobación de una determinada asignatura.

Por lo que, la regresión logística resulta útil para aquellas situaciones en la que se debe predecir la presencia o ausencia de una característica o resultado dado ciertos factores (IBM, 2022), en cierta forma, esta técnica es una hermana distante de la regresión lineal, sin embargo, un modelo basado en regresión logística es ideado para variables dicotómicas.

3.2.2. Árbol de decisión

Un árbol de decisión es “un mapa de los posibles resultados de una serie de decisiones relacionadas” (Lucidchart, s.f.). Sin embargo, dentro del contexto de la ciencia de datos, un “árbol de decisión es un algoritmo de aprendizaje supervisado no paramétrico, que se utiliza tanto para tareas de clasificación como de regresión. Tiene una estructura de árbol jerárquica, que consta de un nodo raíz, ramas, nodos internos y nodos hoja.” (IBM, s.f.).

Adicionalmente, “El aprendizaje del árbol de decisiones emplea una estrategia de divide y vencerás mediante la realización de una búsqueda codiciosa para identificar los puntos de división óptimos dentro de un árbol.” (IBM, s.f.).

3.2.3. *Random Forest*

Ahora, considerando a Random forest, este es un algoritmo de machine learning que “combina la salida de múltiples árboles de decisión para llegar a un único resultado. Su facilidad de uso y flexibilidad han impulsado su adopción, ya que maneja problemas de clasificación y regresión.” (IBM, s.f.). Además, uno de los puntos más fuertes que se resaltan de este algoritmo es su capacidad de controlar el sobreajuste (scikit-learn, s.f.)

3.2.4. *Support Vector Machines (SVM)*

En primer lugar, es necesario conocer que “el objetivo del algoritmo de la máquina de vectores de soporte es encontrar un hiperplano en un espacio N-dimensional (N: el número de características) que clasifique claramente los puntos de datos.” (Gandhi, 2018). Adicionalmente, las ventajas que presente son:

- Eficaz con una gran variedad de dimensiones.
- Es efectivo incluso donde el número de dimensiones es mayor que el de muestras.
- Es versátil ya que se pueden especificar diferentes funciones del Kernel para la función de decisión.

3.2.5. *Redes Neuronales*

“Una red neuronal es un método de la inteligencia artificial que enseña a las computadoras a procesar datos de una manera que está inspirada en la forma en que lo hace el

cerebro humano” (Amazon, s.f.). Una de sus principales ventajas es la de comprender datos estructurados y realizar observaciones generales sin ningún tipo de entrenamiento (Amazon, s.f.).

Generalmente, las redes neuronales son usadas para los siguientes motivos:

- Diagnóstico médico
- Identificación de compuestos
- Identificación de comida
- Predicciones académicas

Y para esta ocasión, también se usará para predicción de rendimiento académico, por lo que resulta un modelo sumamente útil.

3.3. Métricas de evaluación de modelos de machine learning

3.3.1. *Precisión (Accuracy)*

“La precisión se refiere a qué tan cerca está una medida del valor verdadero o aceptado. La precisión se refiere a qué tan cerca están las medidas del mismo artículo entre sí.” (Universidad de Hawaii, s.f.) Es decir, es simplemente el porcentaje de predicciones correctas que hizo el modelo y se calcula como el número total de predicciones correctas dividido por el número total de predicciones.

3.3.2. *Recall o Sensibilidad*

“Es el número de elementos identificados correctamente como positivos del total de positivos verdaderos.” (Universidad de Hawaii, s.f.). Es decir, esta métrica tiene como única labor el responder a la pregunta: "¿Cuántos de los casos positivos reales pude identificar correctamente?".

3.3.3. Precisión (*Precision*)

La métrica de precisión referida a “Precision” en inglés, es más parecida a la métrica recall de lo que es a la métrica de precisión en términos de “Accuracy”, pues, la métrica “Precision” de esta sección responde la pregunta: "Cuando se predice un resultado positivo, ¿cuántas veces se está en lo correcto?" (Universidad de Hawaii, s.f.).

3.3.4. *F1-Score*

“Siempre es mejor establecer como científico de datos una métrica de evaluación de un solo número para que su equipo luego optimice.” (Universidad de Hawaii, s.f.). Esta métrica representa la media armónica de Precision y Recall.

3.3.5. Curva ROC y AUC

ROC es la Curva de Características Operativas del Receptor y AUC es el Área Bajo la Curva, además, “cuando se trata de un problema de clasificación, se puede contar con una curva AUC-ROC. Esta es una de las métricas de evaluación más importante para verificar el rendimiento de cualquier modelo de clasificación.” (aprendeja, s.f.)

CAPÍTULO IV: METODOLOGÍA

4. Metodología

En primera instancia, es necesario recordar al estimado lector que este trabajo de titulación ha sido realizado con la metodología CRISP-DM, esto implica que las fases de esta metodología se relacionan nativamente con el contenido del presente trabajo; por lo que, a partir de esta premisa, se podrá visualizar el desarrollo de las siguientes fases:

- Fase de comprensión de los datos
- Fase de preparación de los datos
- Fase de modelado
- Fase de evaluación.

4.1. Comprensión de los datos

La fase de comprensión de datos es una etapa clave en la metodología CRISP-DM que “implica estudiar más de cerca los datos disponibles de minería. Este paso es esencial para evitar problemas inesperados durante la siguiente fase (preparación de datos).” (IBM, s. f.). Es decir, esta fase es de vital importancia en un proyecto como en el caso de predecir el rendimiento académico universitario de los estudiantes.

Durante esta etapa, el objetivo es obtener una comprensión detallada del conjunto de datos disponible, incluyendo su estructura, calidad y las relaciones entre las variables; lo que significa el acceso a los datos obtenidos y la aplicación métodos de visualización de distribuciones y relaciones para su exploración (IBM, s. f.). En otras palabras, se llevará a cabo una exploración exhaustiva de los conjunto de datos obtenidos para identificar patrones,

relaciones y anomalías que puedan influir en el rendimiento académico de los estudiantes con la finalidad de establecer una hipótesis sobre los patrones y relaciones en los datos.

En resumen, la fase de comprensión de datos es fundamental para el éxito del modelo de predicción y debe ser abordada cuidadosamente para garantizar que el modelo sea preciso y confiable, además, esta etapa sienta las bases para las etapas posteriores del proyecto como la preparación de los datos y la selección del modelo adecuado. Es decir, la fase de comprensión de datos es crucial para desarrollar un modelo predictivo efectivo y confiable, lo que a su vez permite tomar decisiones informadas basadas en los datos.

4.1.1. Librerías requeridas para la fase

En el desarrollo de esta fase se ha recurrido a un conjunto de librerías en Python que han facilitado la manipulación, análisis y visualización de los datos; en esta ocasión se utilizó 'pandas' para la manipulación de datos, ya que 'pandas' es una rápida, poderosa, flexible y fácil de usar herramienta de fuente abierta para el análisis de datos y su manipulación. (pandas, s. f.). Sin embargo, la principal razón de su uso se debe a su rápido y eficiente tipo de objeto llamado DataFrame que es utilizado para la manipulación de datos y cuenta con indexación integrada. (pandas, s. f.). Este objeto llamado 'DataFrame' permitirá la manipulación de los conjuntos de datos obtenidos durante el desarrollo de esta investigación.

Por otra parte, el uso de 'numpy' se enfocó en cálculos específicos, así como permitir la visualización de datos de una manera más específica de lo que únicamente hubiera sido posible con la librería 'pandas'. Con otro enfoque, para la visualización de datos y análisis estadístico, se emplearon varias librerías como: 'matplotlib', 'seaborn' y 'scipy_stats'; de las cuales, en específico, 'Matplotlib' y 'seaborn' permiten generar gráficos de alta calidad, mientras que 'scipy.stats' ofrece funciones estadísticas. Finalmente, se utilizó la librería 'os', que permite

interactuar con el sistema operativo, y como librería adicional se recurrió a 'random', que se emplea para generar números aleatorios. Estas librerías, en conjunto, proporcionan las herramientas necesarias para un manejo eficiente de los datos en el contexto de la ciencia de datos.

Ilustración 1

Librerías requeridas para la fase de comprensión de los datos

Librerías requeridas

```
In [1]: # Librerías para manipular datos
import pandas as pd
import numpy as np

# Librerías para gráficos y estadísticas
import matplotlib.pyplot as plt
import seaborn as sns
import scipy.stats as stats

# Librerías para cargar los datos
import os

# Librerías adicionales
import random
```

4.1.2. Entendimiento de los conjuntos de datos obtenidos

Los datos se obtuvieron tras un acuerdo con la Dirección de Bienestar Estudiantil de la Pontificia Universidad Católica del Ecuador, por lo que, al ser datos reales de la comunidad universitaria se tuvo que firmar un acuerdo de confidencialidad respectivo. De este acuerdo se obtuvieron registros correspondientes a 3 diferentes reportes, los reportes obtenidos fueron:

- Reporte de Asignaturas perdidas por periodo y reincidencia
- Reporte de Calificaciones por carrera

- Reporte de detalle de estudiantes inscritos con beca

A partir de estos reportes, por cada uno se obtuvieron 4 periodos académicos, estos fueron:

- Periodo académico 2021-01
- Periodo académico 2021-02
- Periodo académico 2022-01
- Periodo académico 2021-02

De manera general, todos los reportes correspondían a Pregrado y al campus de la Pontificia Universidad Católica del Ecuador en Quito, sin embargo, en específico, cada reporte consistía en lo siguiente:

- **Reporte de asignaturas perdidas por periodo y reincidencia:** Este reporte contiene registros de los estudiantes que perdieron asignaturas en un determinado periodo, por ejemplo, para el reporte del periodo 2021-01 se contiene información de los estudiantes que perdieron sus asignaturas en el semestre 2021-01.
- **Reporte de calificaciones por carrera:** Este reporte contiene registros de las calificaciones de cada estudiante tanto parciales como finales, así como también se tiene presente las carreras de los estudiantes, asignaturas y profesores.
- **Reporte de detalle de estudiantes inscritos con beca:** Este reporte contiene registros de los estudiantes que poseen becas, discapacidades, cantidad de asignaturas con segunda o tercera matrícula, e información de pago de estos.

4.1.3. Carga de los conjuntos de datos

Con respecto a la carga de los conjuntos de datos obtenidos y comprendidos en la sección anterior, es necesario comprender que estos se importarán como marcos de datos bajo la nomenclatura visible en la siguiente imagen:

Ilustración 2

Nomenclatura de los marcos de datos de los conjuntos de datos obtenidos

Reporte de asignaturas perdidas por periodo y reincidencia:

- Reporte de asignaturas perdidas por periodo y reincidencia 2021-01 -> perdidas01
- Reporte de asignaturas perdidas por periodo y reincidencia 2021-02 -> perdidas02
- Reporte de asignaturas perdidas por periodo y reincidencia 2022-01 -> perdidas03
- Reporte de asignaturas perdidas por periodo y reincidencia 2022-02 -> perdidas04

Reporte de Calificaciones por Carrera:

- Reporte de Calificaciones por Carrera 2021-01 -> calificaciones01
- Reporte de Calificaciones por Carrera 2021-02 -> calificaciones02
- Reporte de Calificaciones por Carrera 2022-01 -> calificaciones03
- Reporte de Calificaciones por Carrera 2022-02 -> calificaciones04

Reporte de detalle de estudiantes inscritos con beca:

- Reporte de detalle de estudiantes inscritos con beca 2021-01 -> beca01
- Reporte de detalle de estudiantes inscritos con beca 2021-02 -> beca02
- Reporte de detalle de estudiantes inscritos con beca 2022-01 -> beca03
- Reporte de detalle de estudiantes inscritos con beca 2022-02 -> beca04

A partir de la nomenclatura visible, es importante mencionar que los diferentes conjuntos de datos que se importarán, además de corresponder a la nomenclatura descrita, estos se recopilaran en listas individuales por reporte, es decir, todos los conjuntos de datos correspondientes a cada reporte serán almacenados en listas específicas de cada reporte. Para los periodos del reporte de asignaturas perdidas se considerará una lista llamada “perdidas”, para los

periodos del reporte de calificaciones por carrera se considerará una lista llamada “calificaciones”, y, finalmente, para los periodos del reporte de becas se considerará una lista llamada “becas”.

Ilustración 3

Función para la importación de conjuntos de datos y colocación de nombres dada una nomenclatura específica

```
In [2]: def get_reportes(directory, reporte_deseado):  
    """  
    La función evaluará un directorio dado y devolverá dataframes en función de los reportes deseados.  
    Si la función no encuentra ningún reporte deseado devolverá una lista vacía.  
    """  
    # Obtenemos la lista de archivos en el directorio  
    files = os.listdir(directory)  
  
    # Creamos una variable que almacenará los reportes como dataframes  
    dataframes = []  
  
    # Se itera en busca de los archivos requeridos  
    for file in files:  
        # Variable con la dirección del archivo evaluado  
        filepath = os.path.join(directory, file)  
        # Evaluamos si es el reporte deseado  
        if os.path.isfile(filepath) and ".join(file.split()[:-1]) == reporte_deseado:  
            # Se extrae el archivo CSV como dataframe  
            temp_df = pd.read_csv(filepath, sep=";", low_memory=False)  
            # Se agrega el dataframe a la lista con todos los dataframes.  
            dataframes.append(temp_df)  
  
    # Se retorna la lista con todos los reportes como dataframes  
    return dataframes
```

Ilustración 4

Código para la importación de los datos en marcos de datos para su uso en el código

```
In [3]: # Variables de directorio y reportes deseados
directorio = "../Datos Estudiantes/"
reportes_deseados = [
    "Reporte de asignaturas perdidas por periodo y reincidencia",
    "Reporte de Calificaciones por Carrera",
    "Reporte de detalle de estudiantes inscritos con beca"
]

# Extraemos los reportes de asignaturas perdidas
perdidas = get_reportes(directorio, reportes_deseados[0])
# Extraemos los reportes de calificaciones
calificaciones = get_reportes(directorio, reportes_deseados[1])
# Extraemos los reportes de becas
becas = get_reportes(directorio, reportes_deseados[2])

# Damos nombres a los dataframes
for i in range(4):
    perdidas[i].name = "perdidas0" + str(i+1)
    calificaciones[i].name = "calificaciones0" + str(i+1)
    becas[i].name = "becas0" + str(i+1)

# Verificación de nombres
print("Los reportes importados son:")
print("\nReportes de asignaturas perdidas: ")
for perdida in perdidas: print(f"\t- {perdida.name}")
print("\nReportes de calificaciones por carrera: ")
for calificacion in calificaciones: print(f"\t- {calificacion.name}")
print("\nReportes de becas: ")
for beca in becas: print(f"\t- {beca.name}")
```

Ilustración 5

Resultado de la importación de los conjuntos de datos

```
Los reportes importados son:

Reportes de asignaturas perdidas:
- perdidas01
- perdidas02
- perdidas03
- perdidas04

Reportes de calificaciones por carrera:
- calificaciones01
- calificaciones02
- calificaciones03
- calificaciones04

Reportes de becas:
- becas01
- becas02
- becas03
- becas04
```

Los códigos presentados dieron como resultado la importación exitosa de los conjuntos de datos compartidos por la dirección de estudiantes de la Pontificia Universidad Católica del Ecuador, es decir, los conjuntos de datos explicados en la sección anterior se encuentran listos para su uso en los siguientes capítulos de la fase de Comprensión de los Datos.

Sin embargo, antes de continuar con la verificación de la consistencia, es prudente divulgar las dimensiones que cada marco de datos posee, para esto, se realizó un código capaz de mostrar las dimensiones de los conjuntos de datos importados, es decir, se permitirá visualizar la cantidad de variables y registros que poseen los diferentes periodos de cada reporte; no obstante, es importante destacar que “una variable es una característica de los elementos u objetos que se estudian. Y los datos son los valores que se obtienen para cada variable” (Agacino, s. f.). Donde

la unión de los datos de todas las variables conforma un registro, los siguientes datos de todas las variables corresponden a un segundo registro, y así sucesivamente.

El código para la impresión de las dimensiones mencionadas y sus resultados, se pueden encontrar a continuación:

Ilustración 6

Código para la impresión de las dimensiones de los conjuntos de datos importados

```
In [4]: # Se genera una lista con todos los dataframes para facilitar el proceso de impresión
reportes = []
reportes.extend(perdidas)
reportes.extend(calificaciones)
reportes.extend(becas)

# Loop para la impresión de todos los conjuntos de datos
for dataframe in reportes:
    # Se imprime un encabezado para cada uno de los reportes
    if dataframe.name == "perdidas01":
        print("\n" + "*" * 100 + "\n")
        print("Analizando los reportes de Asignaturas perdidas:\n")
    elif dataframe.name == "calificaciones01":
        print("\n" + "*" * 100 + "\n")
        print("Analizando los reportes de Calificaciones por Carrera:\n")
    elif dataframe.name == "becas01":
        print("\n" + "*" * 100 + "\n")
        print("Analizando los reportes de Becas:\n")

    # Impresión de las dimensiones
    print(f"Analizando el reporte {dataframe.name}:")
    print(f"\t- El reporte posee {dataframe.shape[0]} registros.")
    print(f"\t- El reporte posee {dataframe.shape[1]} variables.\n")
```

El código presente permite visualizar la manera en la que se desplegarán las diferentes dimensiones de los periodos de los tres reportes, para esto, es necesario recordar la definición de lo que son los registros y variables, que, como se mencionó previamente, las variables son características de los elementos u objetos de estudio, en esta ocasión, de los estudiantes hallados en los diferentes reportes; mientras que, los registros son los datos que corresponden a cada estudiante con respecto a las características que se dispone en dichos reportes.

Ilustración 7

Dimensiones de los conjuntos de datos del reporte de Asignaturas Perdidas

```
*****  
  
Analizando los reportes de Asignaturas perdidas:  
  
Analizando el reporte perdidas01:  
  - El reporte posee 1670 registros.  
  - El reporte posee 23 variables.  
  
Analizando el reporte perdidas02:  
  - El reporte posee 2104 registros.  
  - El reporte posee 23 variables.  
  
Analizando el reporte perdidas03:  
  - El reporte posee 3069 registros.  
  - El reporte posee 23 variables.  
  
Analizando el reporte perdidas04:  
  - El reporte posee 3141 registros.  
  - El reporte posee 23 variables.
```

En esta ocasión, es visible que, para el reporte de asignaturas perdidas, los registros, es decir, los elementos de estudio poseen 23 características diferentes.

Ilustración 8

Dimensiones de los conjuntos de datos del reporte de Calificaciones por Carrera

```
*****  
  
Analizando los reportes de Calificaciones por Carrera:  
  
Analizando el reporte calificaciones01:  
  - El reporte posee 163810 registros.  
  - El reporte posee 40 variables.  
  
Analizando el reporte calificaciones02:  
  - El reporte posee 158818 registros.  
  - El reporte posee 40 variables.  
  
Analizando el reporte calificaciones03:  
  - El reporte posee 150108 registros.  
  - El reporte posee 40 variables.  
  
Analizando el reporte calificaciones04:  
  - El reporte posee 142205 registros.  
  - El reporte posee 40 variables.
```

Por otra parte, es visible que, para el reporte de calificaciones por carrera, los registros, es decir, los elementos de estudio poseen 40 características diferentes.

Ilustración 9

Dimensiones de los conjuntos de datos del reporte de Becas

```
*****

Analizando los reportes de Becas:

Analizando el reporte becas01:
- El reporte posee 9399 registros.
- El reporte posee 27 variables.

Analizando el reporte becas02:
- El reporte posee 8564 registros.
- El reporte posee 27 variables.

Analizando el reporte becas03:
- El reporte posee 8336 registros.
- El reporte posee 27 variables.

Analizando el reporte becas04:
- El reporte posee 8255 registros.
- El reporte posee 27 variables.
```

Finalmente, es visible que, para el reporte de becas, los registros, es decir, los elementos de estudio poseen 27 características diferentes.

Por lo que, a partir de las ilustraciones visualizadas se puede concluir que se tiene una gran cantidad de datos y variables, por ejemplo, con respecto al reporte de calificaciones por carrera, este posee 40 variables diferentes y un aproximado de 155000 datos; En otras palabras, dada la inmensa cantidad de datos y variables en los distintos reportes, se posee una visión optimista a priori de los futuros resultados de este trabajo de titulación.

4.1.4. Verificación de consistencia entre periodos de los conjuntos de datos

En esta sección se realizará una verificación de que los datos entre periodos sean consistentes por tipo de reporte, esto es importante debido a que permitirá realizar la identificación de variables y otros procesos de manera más sencilla y reducida. Es decir, se

realizará una verificación de que las variables que cada reporte posee sean consistentes independiente del periodo. Para esto, nuevamente se realizó un código, donde, para empezar, se definieron dos funciones que permitirán no solamente un análisis prudente de la consistencia, sino que permitirán visualizar los resultados de una manera elegante.

Ilustración 10

Funciones generales para el análisis de la consistencia entre periodos de un reporte

```
In [5]: def cute_print(dataframe):  
    '''  
    Esta función imprime de manera bonita las columnas de los dataframes.  
    '''  
    columns = dataframe.columns  
    print("Las columnas son:", end=" ")  
    for column in columns: print(column, end=" | ")  
    print(f"\n\nExisten en total: {len(columns)} columnas")  
  
def compare_columns(dataframes):  
    '''  
    Esta función evalúa si las columnas de los dataframes ingresados son iguales.  
    Devuelve "Sí" o "No" correspondientemente.  
    '''  
    # Creamos una lista con todas las columnas  
    columns = [list(dataframe.columns) for dataframe in dataframes]  
  
    # Realizamos la comparación entre las columnas de los dataframes ingresados  
    comparativa = "Sí" if all(columns[i] == columns[i+1] for i in range(len(columns)-1)) else "No"  
  
    # Se imprime el resultado de la comparativa  
    print(f"Las columnas de los dataframes son iguales? {comparativa}")
```

A partir de estas dos funciones, se realizó la verificación de la consistencia de los periodos de los tres diferentes reportes, para ello, se llevó a cabo el siguiente código para analizar el reporte de asignaturas perdidas:

Ilustración 11

Código para el análisis de consistencia del reporte de Asignaturas Perdidas

```
'''
Imprime las variables que posee el reporte y su cantidad
'''

for perdida in perdidas:
    print("\n" + "*" * 100)
    print(f"\nAnalizando {perdida.name}...\n")
    cute_print(perdida)

print("\n" + "*" * 100)
print("\nAnalizando la consistencia de las variables...\n")
compare_columns(perdidas)
```

Como se puede observar, el código llama a las funciones previamente definidas para esta sección, de esta forma, los resultados de este pueden ser apreciados a continuación:

Ilustración 12

Parte 1 de los resultado del código para el análisis de las variables y su cantidad del reporte de Asignaturas Perdidas (corresponden al primer y segundo periodo del reporte)

```
*****

Analizando perdidas01...

Las columnas son: CEDULA | IDD | PIDM | APELLIDO | SEGUNDO APELLIDO | NOMBRE | PERIODO | PLAN_E | CAMP | DESC_FACULTAD | PROGRAMA
| DESC_PROGRAMA | NIVEL_ESTUDIANTE | NF | NRC | MATE | CRSE | SHRTCKN_CRSE_TITLE | NOTA_FINAL | APROBADO | VECES | MODO | CRED |

Existen en total: 23 columnas

*****

Analizando perdidas02...

Las columnas son: CEDULA | IDD | PIDM | APELLIDO | SEGUNDO APELLIDO | NOMBRE | PERIODO | PLAN_E | CAMP | DESC_FACULTAD | PROGRAMA
| DESC_PROGRAMA | NIVEL_ESTUDIANTE | NF | NRC | MATE | CRSE | SHRTCKN_CRSE_TITLE | NOTA_FINAL | APROBADO | VECES | MODO | CRED |

Existen en total: 23 columnas
```

Ilustración 13

Parte 2 de los resultado del código para el análisis de las variables y su cantidad del reporte de Asignaturas Perdidas (corresponden al tercer y cuarto periodo del reporte)

```
*****

Analizando perdidas03...

Las columnas son: CEDULA | IDD | PIDM | APELLIDO | SEGUNDO APELLIDO | NOMBRE | PERIODO | PLAN_E | CAMP | DESC_FACULTAD | PROGRAMA
| DESC_PROGRAMA | NIVEL_ESTUDIANTE | NF | NRC | MATE | CRSE | SHRTCKN_CRSE_TITLE | NOTA_FINAL | APROBADO | VECES | MODO | CRED |

Existen en total: 23 columnas

*****

Analizando perdidas04...

Las columnas son: CEDULA | IDD | PIDM | APELLIDO | SEGUNDO APELLIDO | NOMBRE | PERIODO | PLAN_E | CAMP | DESC_FACULTAD | PROGRAMA
| DESC_PROGRAMA | NIVEL_ESTUDIANTE | NF | NRC | MATE | CRSE | SHRTCKN_CRSE_TITLE | NOTA_FINAL | APROBADO | VECES | MODO | CRED |

Existen en total: 23 columnas
```

Ilustración 14

Parte Final de los resultado del código para el análisis de las variables y su cantidad del reporte de Asignaturas Perdidas (Resultado de la consistencia de las variables)

```
Analizando la consistencia de las variables...

Las columnas de los dataframes son iguales? Sí
```

Como se puede observar, dado los resultados del código realizado, se pudo determinar que no existen inconsistencias entre los diferentes periodos del reporte de asignaturas perdidas. A continuación, se define el código para realizar el análisis del reporte de calificaciones por carrera:

Ilustración 15

Código para el análisis de consistencia del reporte de Calificaciones por Carrera

```
: '''
Imprime las variables que posee el reporte y su cantidad
'''

for calificacion in calificaciones:
    print("\n" + "*" * 100)
    print(f"\nAnalizando {calificacion.name}...\n")
    cute_print(calificacion)

print("\n" + "*" * 100)
print("\nAnalizando la consistencia de las variables...\n")
compare_columns(calificaciones)
```

Similarmente, se puede observar el empleo de las dos funciones declaradas anteriormente y necesarias para el análisis de la consistencia, por esta razón, los resultados del código se podrán presenciar a continuación:

Ilustración 16

Parte 1 de los resultado del código para el análisis de las variables y su cantidad del reporte de Calificaciones por Carrera (corresponden al primer y segundo periodo del reporte)

```
*****

Analizando calificaciones01...

Las columnas son: pidm_estudiante | id_banner_estudiante | cedula_pasaporte_estudiante | apellidos_estudiante | nombres_estudiante |
codigo_periodo | desc_periodo | codigo_facultad | facultad | codigo_carrera | carrera | numero_curso | titulo_curso | NRC | numero_pa
ralelo | nombre_parcial | descripcion_parcial | peso | nota_maxima | nota_parcial | porcentaje_nota_parcial | nota_final | pidm_docen
te | id_banner_docente | cedula_pasaporte_docente | apellidos_docente | nombres_docente | codigo_materia | descripcion_materia | secu
encial_parcial | fecha_nota_parcial | fecha_nota_final | codigo_programa | descripcion_programa | nota_final_SHATCKG | cod_razon_de_c
ambio | desc_razon_de_cambio | calificacion_del_parcial_score | codigo_comentario_nota | desc_comentario_nota |

Existen en total: 40 columnas

*****

Analizando calificaciones02...

Las columnas son: pidm_estudiante | id_banner_estudiante | cedula_pasaporte_estudiante | apellidos_estudiante | nombres_estudiante |
codigo_periodo | desc_periodo | codigo_facultad | facultad | codigo_carrera | carrera | numero_curso | titulo_curso | NRC | numero_pa
ralelo | nombre_parcial | descripcion_parcial | peso | nota_maxima | nota_parcial | porcentaje_nota_parcial | nota_final | pidm_docen
te | id_banner_docente | cedula_pasaporte_docente | apellidos_docente | nombres_docente | codigo_materia | descripcion_materia | secu
encial_parcial | fecha_nota_parcial | fecha_nota_final | codigo_programa | descripcion_programa | nota_final_SHATCKG | cod_razon_de_c
ambio | desc_razon_de_cambio | calificacion_del_parcial_score | codigo_comentario_nota | desc_comentario_nota |

Existen en total: 40 columnas
```

Ilustración 17

Parte 2 de los resultado del código para el análisis de las variables y su cantidad del reporte de Calificaciones por Carrera (corresponden al tercer y cuarto periodo del reporte)

```
*****

Analizando calificaciones03...

Las columnas son: pidm_estudiante | id_banner_estudiante | cedula_pasaporte_estudiante | apellidos_estudiante | nombres_estudiante |
codigo_perodo | desc_perodo | codigo_facultad | facultad | codigo_carrera | carrera | numero_curso | titulo_curso | NRC | numero_pa
ralelo | nombre_parcial | descripcion_parcial | peso | nota_maxima | nota_parcial | porcentaje_notaparcial | nota_final | pidm_docen
te | id_banner_docente | cedula_pasaporte_docente | apellidos_docente | nombres_docente | codigo_materia | descripcion_materia | secu
encial_parcial | fecha_notaparcial | fecha_notafinal | codigo_programa | descripcion_programa | nota_final_SHATCKG | cod_razon_de_c
ambio | desc_razon_de_cambio | calificacion_del_parcial_score | codigo_comentario_nota | desc_comentario_nota |

Existen en total: 40 columnas

*****

Analizando calificaciones04...

Las columnas son: pidm_estudiante | id_banner_estudiante | cedula_pasaporte_estudiante | apellidos_estudiante | nombres_estudiante |
codigo_perodo | desc_perodo | codigo_facultad | facultad | codigo_carrera | carrera | numero_curso | titulo_curso | NRC | numero_pa
ralelo | nombre_parcial | descripcion_parcial | peso | nota_maxima | nota_parcial | porcentaje_notaparcial | nota_final | pidm_docen
te | id_banner_docente | cedula_pasaporte_docente | apellidos_docente | nombres_docente | codigo_materia | descripcion_materia | secu
encial_parcial | fecha_notaparcial | fecha_notafinal | codigo_programa | descripcion_programa | nota_final_SHATCKG | cod_razon_de_c
ambio | desc_razon_de_cambio | calificacion_del_parcial_score | codigo_comentario_nota | desc_comentario_nota |

Existen en total: 40 columnas
```

Ilustración 18

Parte Final de los resultado del código para el análisis de las variables y su cantidad del reporte de Calificaciones por Carrera (Resultado de la consistencia de las variables)

```
Analizando la consistencia de las variables...

Las columnas de los dataframes son iguales? Sí
```

En esta ocasión, nuevamente es observable que los resultados del código realizado permitieron determinar que no existen inconsistencias en lo que respecta al tipo de variables que contienen los diferentes periodos del reporte de calificaciones por carrera, lo cual es un indicador positivo. Finalmente, a continuación, se define el código que se empelará para realizar el análisis de consistencia del reporte de becas:

Ilustración 19

Código para el análisis de consistencia de los periodos del reporte de Becas

```
'''
Imprime las variables que posee el reporte y su cantidad
'''

for beca in becas:
    print("\n" + "*" * 100)
    print(f"\nAnalizando {beca.name}...\n")
    cute_print(beca)

print("\n" + "*" * 100)
print("\nAnalizando la consistencia de las variables...\n")
compare_columns(becas)
```

De manera similar a como se ha realizado para los reportes de asignaturas perdidas y calificaciones por carrera, en esta ocasión, se hace uso de las funciones “cute_print” y “compare_columns” para el análisis de la consistencia de los periodos del reporte de becas. Los resultados de este código podrán ser vistos a continuación:

Ilustración 20

Parte 2 de los resultado del código para el análisis de las variables y su cantidad del reporte de Becas (corresponden al primer y segundo periodo del reporte)

```
*****

Analizando becas01...

Las columnas son: PIDM | ID_BANNER | APELIDOS | PRIMER_NOMBRE | SEGUNDO_NOMBRE | CEDULA | NACIONALIDAD | CELULAR | TELEFONO | CORREO_
PERSONAL | CORREO_INSTITUC | GENERO | ETNIA | CODIGO_COLEGIO | TIPO_COLEGIO | PROGRAMA | PROGAMA_DESC | CAMPUS_PROGRAMA | AREA_MATERI
AS | NIVEL_POR_MATERIAS | NIVEL_POR_ASIG_APROBADAS | CATEGORIA_BECA_SOCIOECONOMICA | TIPO_REGIMEN | NUM_MATERIAS_SEGUNDA_MATRICULA |
NUM_MATERIAS_TERCERA_MATRICULA | TIENE_PAGO | TIENE_SALDO |

Existen en total: 27 columnas

*****

Analizando becas02...

Las columnas son: PIDM | ID_BANNER | APELIDOS | PRIMER_NOMBRE | SEGUNDO_NOMBRE | CEDULA | NACIONALIDAD | CELULAR | TELEFONO | CORREO_
PERSONAL | CORREO_INSTITUC | GENERO | ETNIA | CODIGO_COLEGIO | TIPO_COLEGIO | PROGRAMA | PROGAMA_DESC | CAMPUS_PROGRAMA | AREA_MATERI
AS | NIVEL_POR_MATERIAS | NIVEL_POR_ASIG_APROBADAS | CATEGORIA_BECA_SOCIOECONOMICA | TIPO_REGIMEN | NUM_MATERIAS_SEGUNDA_MATRICULA |
NUM_MATERIAS_TERCERA_MATRICULA | TIENE_PAGO | TIENE_SALDO |

Existen en total: 27 columnas
```


Ilustración 21

Parte 2 de los resultado del código para el análisis de las variables y su cantidad del reporte de Becas (corresponden al tercer y cuarto periodo del reporte)

```
*****
Analizando becas03...

Las columnas son: PIDM | ID_BANNER | APELIDOS | PRIMER_NOMBRE | SEGUNDO_NOMBRE | CEDULA | NACIONALIDAD | CELULAR | TELEFONO | CORREO_
PERSONAL | CORREO_INSTITUC | GENERO | ETNIA | CODIGO_COLEGIO | TIPO_COLEGIO | PROGRAMA | PROGAMA_DESC | CAMPUS_PROGRAMA | AREA_MATERI
AS | NIVEL_POR_MATERIAS | NIVEL_POR_ASIG_APROBADAS | CATEGORIA_BECA_SOCIOECONOMICA | TIPO_REGIMEN | NUM_MATERIAS_SEGUNDA_MATRICULA |
NUM_MATERIAS_TERCERA_MATRICULA | TIENE_PAGO | TIENE_SALDO |

Existen en total: 27 columnas

*****
Analizando becas04...

Las columnas son: PIDM | ID_BANNER | APELIDOS | PRIMER_NOMBRE | SEGUNDO_NOMBRE | CEDULA | NACIONALIDAD | CELULAR | TELEFONO | CORREO_
PERSONAL | CORREO_INSTITUC | GENERO | ETNIA | CODIGO_COLEGIO | TIPO_COLEGIO | PROGRAMA | PROGAMA_DESC | CAMPUS_PROGRAMA | AREA_MATERI
AS | NIVEL_POR_MATERIAS | NIVEL_POR_ASIG_APROBADAS | CATEGORIA_BECA_SOCIOECONOMICA | TIPO_REGIMEN | NUM_MATERIAS_SEGUNDA_MATRICULA |
NUM_MATERIAS_TERCERA_MATRICULA | TIENE_PAGO | TIENE_SALDO |

Existen en total: 27 columnas
```

Ilustración 22

Parte Final de los resultado del código para el análisis de las variables y su cantidad del reporte de Becas (Resultado de la consistencia de las variables)

```
Analizando la consistencia de las variables...

Las columnas de los dataframes son iguales? Sí
```

En síntesis, se pudo concluir que tampoco existen inconsistencias con lo que respecta al tipo de variables que ostentan los diferentes periodos del reporte de becas, es decir, a partir de las conclusiones presentadas, es factible afirmar que en los tres diferentes reportes analizados no se encontraron inconsistencias entre las características de cada periodo.

4.1.5. Identificación de las variables de los conjuntos de datos

La identificación de las variables en los conjuntos de datos es un paso esencial en cualquier estudio académico que implica un análisis de datos, ya que permite al investigador

conocer el sentido de las variables más allá de los datos registrados en ellas, es decir, es una introducción a los datos (UNAM, s. f.). Por lo que, esta etapa permite entender la estructura y la naturaleza de los datos disponibles, esto representa un factor crítico en la determinación de las técnicas de análisis apropiadas a utilizar.

No obstante, es importante mencionar que en el proceso de la identificación de las características que están presentes en cada reporte, se considerará todos los periodos de cada uno como uno solo, es decir, se considerará las variables de un solo periodo de todos los periodos de cada reporte; esto es posible debido a que en la sección anterior se estableció que no hay inconsistencias en las variables entre periodos por cada reporte.

Por lo tanto, para comenzar, las variables o características disponibles en el reporte de asignaturas perdidas son las siguientes:

Tabla 1

Identificación de Variables del Reporte de Asignaturas Perdidas

Variable	Descripción
CEDULA	Número de cédula del estudiante.
IDD	Identificador único del estudiante en el banner.
PIDM	Identificador asociado al registro.
APELLIDO	Primer apellido del estudiante.
SEGUNDO_APELLIDO	Segundo apellido del estudiante.
NOMBRE	Nombre del estudiante.
PERIODO	Periodo académico en el que se perdió la asignatura.
PLAN_E	Código del plan de estudios al que pertenece el estudiante.

CAMP	Campus al que pertenece el estudiante.
DESC_FACULTAD	Nombre completo de la facultad a la que pertenece el estudiante.
PROGRAMA	Código del programa académico al que pertenece el estudiante.
DESC_PROGRAMA	Nombre del programa en el que se encuentra inscrito el estudiante.
NIVEL_ESTUDIANTE	Nivel académico del estudiante dentro del programa.
NF	Nivel académico en el que se encuentra el estudiante, puede ser Grado, Idiomas, Técnico Superior, etc.
NRC	Número de referencia de registro de curso, identificador único para el curso al momento de registrarlo.
MATE	Código del área a la que pertenece la asignatura.
CRSE	Número de la asignatura.
SHRTCKN_CRSE_TITLE	Nombre de la asignatura perdida.
NOTA_FINAL	Calificación final del estudiante en el curso.
APROBADO	Indicador si el estudiante aprobó o reprobó el curso.
VECES	Número de veces que el estudiante ha tomado el curso.
MODO	Escala de calificación, puede ser G para General.
CRED	Número de créditos asignados al curso.

A partir de la identificación de las variables del reporte de asignaturas perdidas, es importante notar que los valores de MATE y NF se encuentran en código, sus valores serán descritos en las siguientes tablas:

Tabla 2

Diccionario de Códigos MATE (Áreas de Asignaturas)

Código	Descripción
1	Educación
2	Artes y Humanidades
3	Cien.Soc.period.form y Derecho
4	Administración
5	Cien.Natur.Matemát.y Estadíst.
6	Tecno.Informac.y Comun.(TIC)
7	Ingeniería Industria y Constr.
8	Agricul.Silvic.pesca y veteri.
9	Salud y Bienestar
10	Servicios
11	Investigación

Tabla 3

Diccionario de Códigos de NF (Nivel Académico)

Código	Descripción
10	Educación continua
20	Idiomas

54	Técnico Superior
55	Tecnológico Superior
56	Tecnológico Superior Universitario
65	Grado
70	Diplomado
73	Especialización
74	Maestría Investigación
75	Maestría Académica
76	Maestría Académica Nacional
77	Especialización Nacional
84	Doctorado (PhD)

Hasta este momento se han identificado todas las variables existentes con sus significados y sentidos del reporte de asignaturas perdidas, a continuación, se identificarán las variables del reporte de calificaciones por carrera:

Tabla 4
Identificación de Variables del Reporte de Calificaciones por Carrera

Variable	Descripción
pidm_estudiante	Identificador asociado al registro.
id_banner_estudiante	Identificador único del estudiante en el sistema Banner.

cedula_pasaporte_estudiante	Número de cédula o pasaporte del estudiante.
apellidos_estudiante	Apellidos del estudiante.
nombres_estudiante	Nombres del estudiante.
codigo_periodo	Código del período académico en el que se registraron las calificaciones.
desc_periodo	Nombre del periodo académico en el que se registraron las calificaciones.
codigo_facultad	Código de la facultad a la que pertenece el estudiante.
facultad	Nombre de la facultad a la que pertenece el estudiante.
codigo_carrera	Código de la carrera a la que pertenece el estudiante.
carrera	Nombre de la carrera a la que pertenece el estudiante.
numero_curso	Número del curso específico que el estudiante tomo.
titulo_curso	Nombre del curso específico que el estudiante tomo.
NRC	Número de referencia de registro de curso, identificador único para el curso al momento de registrarlo.
numero_paralelo	Número de paralelo en la que el estudiante tomó el curso.
nombre_parcial	Nombre del parcial del que se tomó el registro de la calificación del estudiante.
descripcion_parcial	Descripción del parcial del que se tomó el registro de la calificación del estudiante.
peso	Peso del parcial en la calificación final del curso.

nota_maxima	Calificación máxima posible en el parcial.
nota_parcial	Calificación que el estudiante obtuvo en el parcial.
porcentaje_nota_parcial	Porcentaje de la nota parcial con respecto a la nota máxima.
nota_final	Calificación final del estudiante en el curso.
pidm_docente	Identificador único del docente asociado al registro.
id_banner_docente	Identificador único del docente en el sistema de gestión académica.
cedula_pasaporte_docente	Número de cédula o pasaporte del docente.
apellidos_docente	Apellidos del docente.
nombres_docente	Nombres del docente.
codigo_materia	Código de la materia cursada.
descripcion_materia	Área de la materia cursada.
secuencial_parcial	Número secuencial asignado a la evaluación específica (por ejemplo, primer parcial (1), segundo parcial (2)).
fecha_nota_parcial	Fecha en la que se registró la calificación de la evaluación específica.
fecha_nota_final	Fecha y hora en la que se registró la nota final del curso.
codigo_programa	Código del programa en el que se encuentra inscrito el estudiante.
descripcion_programa	Nombre del programa en el que se encuentra inscrito el estudiante.

nota_final_SHATCKG	Nota final según el registro oficial de calificaciones.
cod_razon_de_cambio	Código de la razón del cambio de la calificación.
desc_razon_de_cambio	Descripción de la razón del cambio de la calificación.
calificacion_del_parcial_score	Calificación máxima de una evaluación específica.
codigo_comentario_nota	Código que representa algún comentario relacionado con la calificación del estudiante.
desc_comentario_nota	Descripción detallada del comentario relacionado con la calificación del estudiante.

En esta ocasión, el reporte de calificaciones por carrera no cuenta con variables en código por lo que no se necesitan tablas adicionales para identificar el sentido y significado de sus variables. Finalmente, a continuación, se identificarán las variables del reporte de becas:

Tabla 5
Identificación de variables del reporte de Becas

Variable	Descripción
PIDM	Identificador asociado al registro.
ID_BANNER	Identificador único del estudiante en el sistema Banner.
APELIDOS	Apellidos del estudiante.
PRIMER_NOMBRE	Primer nombre del estudiante.
SEGUNDO_NOMBRE	Segundo nombre del estudiante.
CEDULA	Número de cédula del estudiante.

NACIONALIDAD	Nacionalidad del estudiante.
TELEFONO	Número de teléfono del estudiante.
CELULAR	Número de teléfono celular del estudiante.
CORREO_PERSONAL	Correo electrónico personal del estudiante.
CORREO_INSTITUC	Correo electrónico institucional del estudiante.
GENERO	Género del estudiante.
ETNIA	Etnia a la que pertenece el estudiante.
CODIGO_COLEGIO	Nombre del colegio donde el estudiante completó su educación secundaria.
TIPO_COLEGIO	Tipo de colegio donde el estudiante completó su educación secundaria.
PROGRAMA	Código de la carrera a la que pertenece el estudiante.
PROGAMA_DESC	Nombre de la carrera a la que pertenece el estudiante.
CAMPUS_PROGRAMA	Campus en el que se encuentra el estudiante.
AREA_MATERIAS	Identificador del área y curso durante ese periodo.
NIVEL_POR_MATERIAS	Nivel del estudiante.
NIVEL_POR_ASIG_APROBADAS	Nivel de la asignatura aprobada.
CATEGORIA_BECA_SOCIOECONOMICA	Categoría de la beca socioeconómica del estudiante.
TIPO_REGIMEN	Tipo de política de asignación de beca socioeconómica.

NUM_MATERIAS_SEGUNDA_MATRICULA	Número de materias en la segunda matrícula del estudiante.
NUM_MATERIAS_TERCERA_MATRICULA	Número de materias en la tercera matrícula del estudiante.
TIENE_PAGO	Indica si el estudiante tiene pago registrado.
TIENE_SALDO	Indica si el estudiante tiene saldo pendiente.

En conclusión, se ha podido identificar todas las variables que forman parte de los distintos reportes independiente de sus periodos, además, es evidente que existen varias variables que se relacionan entre reportes como, por ejemplo, los identificadores de los estudiantes. Por otra parte, también es visible que existen varias variables que representan información personal identificable, como, por ejemplo, la variable CELULAR del reporte de becas; es necesario realizar un proceso de tratamiento de estas variables que presentan información personal identificable para preservar la confidencialidad y seguridad de los datos de los estudiantes y docentes de la Pontificia Universidad Católica del Ecuador.

4.1.6. Identificación de la información personal identificable

La identificación de la información personal identificable es un aspecto crítico en la manipulación y análisis de datos, debido a que como la constitución de la república del Ecuador establece: “Se reconoce y garantizará a las personas: (...) 19. El derecho a la protección de datos de carácter personal, que incluye el acceso y la decisión sobre información y datos de este carácter, así como su correspondiente protección.” (artículo 66, 2008). Es decir, dado las preocupaciones éticas y legales que rodean a la información personal de los estudiantes y

docente, y que se vive actualmente en un mundo cada vez más digitalizado, la privacidad de los datos personales se ha convertido en una cuestión de importancia primordial (Banco Mundial, 2020); además, la identificación y tratamiento de la información personal identificable es fundamental para garantizar que se cumpla con las leyes y regulaciones de protección de datos, como el Reglamento General de Protección de Datos (GDPR) en Europa o la reciente ley de protección de datos del Ecuador, cuyo objetivo es el “garantizar el derecho que tienen todos los ciudadanos ecuatorianos a que se resguarden sus datos personales, a poder acceder libremente a dicha información y a decidir sobre ella”. (Banco Pichincha, 2022)

Por lo tanto, esta sección, se enfocará en la identificación de información personal identificable (PII por sus siglas en inglés), este tipo de información se refiere a cualquier dato que pueda usarse para identificar a una persona de manera directa o indirecta (Google, s.f.), es decir, su información personal; sin embargo, es necesario comprender que la información personal no es información o un conjunto de información que no se relaciona con la identificación de un individuo. (United Kingdom. Information Commissioner's Office, 2017, pág. 11). Es decir, la información personal es aquella que puede ser usada para la identificación de un individuo, en esta ocasión de estudiantes y docentes de la Pontificia Universidad Católica del Ecuador.

Para lograr la identificación de estos identificadores de información personal se desarrolló un archivo de Excel llamado “Matriz de evaluación de Información Personal Identificable.xlsx” que puede ser encontrado entre los anexos disponibles de este trabajo de titulación, en este documento se analizaron todas las variables disponibles en cada reporte y se determinó que tratamiento se debería ejercer sobre estas variables; sin embargo, es necesario mencionar que este proceso fue llevado a cabo siguiendo los lineamientos de lo que considera

Google como información personal identificable., que menciona que “Google considera información personal identificable la información que se pueda usar por sí sola para identificar o ubicar con precisión a una persona, o para ponerse en contacto con ella de forma directa.” (Google, s.f.). En específico, Google considera como información personal a las “direcciones de correo electrónico, direcciones de correo postal, números de teléfono, ubicaciones precisas y nombres completos (nombre y apellidos).” (Google, s.f.).

Teniendo lo mencionado presente, a continuación, se presentarán las conclusiones de la matriz de evaluación de información personal identificable en conjunto con un método de procesamiento de las variables consideradas.

Como punto de partida, se podrá visualizar las conclusiones del reporte de asignaturas perdidas:

Tabla 6
Identificación de PIIs del Reporte de Asignaturas Perdidas

Variable	¿Por qué es un PII?	Método de procesamiento
CEDULA	Se refiere a un código de identificación único del estudiante.	Supresión
IDD	Se refiere a un código de identificación único del estudiante.	Enmascaramiento
PIDM	Se refiere a un código de identificación único del estudiante.	Supresión
APELLIDO	Se refiere a nombres completos (nombre y apellidos)	Supresión
SEGUNDO_APELLIDO	Se refiere a nombres completos (nombre y apellidos)	Supresión
NOMBRE	Se refiere a nombres completos (nombre y apellidos)	Supresión

NRC	Al ser un identificador vinculado al registro de una asignatura en un periodo determinado, refiere a un código de identificación único	Supresión
-----	--	-----------

En segundo lugar, es el turno de presentar las conclusiones del reporte de calificaciones por carrera:

Tabla 7
Identificación de PII's del Reporte de Calificaciones por Carrera

Variable	¿Por qué es un PII?	Método de procesamiento
pidm_estudiante	Se refiere a un código de identificación único del estudiante.	Supresión
id_banner_estudiante	Se refiere a un código de identificación único del estudiante.	Enmascaramiento
cedula_pasaporte_estudiante	Se refiere a un código de identificación único del estudiante.	Supresión
apellidos_estudiante	Se refiere a nombres completos (nombre y apellidos)	Supresión
nombres_estudiante	Se refiere a nombres completos (nombre y apellidos)	Supresión
NRC	Al ser un identificador vinculado al registro de una asignatura en un periodo determinado, refiere a un código de identificación único	Supresión

pidm_docente	Se refiere a un código de identificación único del docente.	Supresión
id_banner_docente	Se refiere a un código de identificación único del docente.	Enmascaramiento
cedula_pasaporte_docente	Se refiere a un código de identificación único del docente.	Supresión
apellidos_docente	Se refiere a nombres completos (nombre y apellidos)	Supresión
nombres_docente	Se refiere a nombres completos (nombre y apellidos)	Supresión
codigo_comentario_nota	La divulgación de esta variable podría revelar información confidencial relacionada con la calificación del estudiante.	Supresión
desc_comentario_nota	La divulgación de esta variable podría revelar información confidencial relacionada con la calificación del estudiante.	Supresión

Finalmente, es el momento de presentar las conclusiones del reporte de becas:

Tabla 8
Identificación de PIIs del Reporte de Becas

Variable	¿Por qué que es un PII?	Método de procesamiento
----------	-------------------------	-------------------------

PIDM	Se refiere a un código de identificación único del estudiante.	Supresión
ID_BANNER	Se refiere a un código de identificación único del estudiante.	Enmascaramiento
APELLIDOS	Se refiere a nombres completos (nombre y apellidos)	Supresión
PRIMER_NOMBRE	Se refiere a nombres completos (nombre y apellidos)	Supresión
SEGUNDO_NOMBRE	Se refiere a nombres completos (nombre y apellidos)	Supresión
CEDULA	Se refiere a un código de identificación único del estudiante.	Supresión
TELEFONO	Se refiere a un número de teléfono	Supresión
CELULAR	Se refiere a un número de celular	Supresión
CORREO_PERSONAL	Se refiere a un correo electrónico	Supresión
CORREO_INSTITUC	Se refiere a un correo electrónico	Supresión
CODIGO_COLEGIO	Se refiere a un código de identificación único del colegio del estudiante.	Supresión

En conclusión, dada las conclusiones de la matriz de evaluación de información personal identificable, se pudo identificar que existe una gran cantidad de variables que podrían exponer tanto a estudiantes como docentes a problemas relacionados a la seguridad de sus datos, es por ello que, se presenta la necesidad de ejecutar los métodos de procesamiento considerados para tratar estas variables y, de esta forma, preservar la integridad, seguridad y confidencialidad de los datos de los estudiantes y docentes.

4.1.7. Anonimización de datos

A partir de las conclusiones de la sección anterior, se pudieron identificar varias variables que podrían poner en riesgo la seguridad, integridad y confidencialidad de los datos de los estudiantes y docentes de la Pontificia Universidad Católica del Ecuador; por lo que es necesario realizar un proceso llamado “anonimización de datos”, el cual consiste en aplicar técnicas de procesamiento que permite que las variables con datos asociados a información personal identificable cesen de serlo. (Unión Europea, 2016). En consecuencia, a partir igualmente de las conclusiones de la matriz presentada en la sección anterior, se pudo identificar que técnicas serían las más apropiadas para cada caso.

Sin embargo, es importante mencionar que en específico la anonimización, como se menciona en el 2017, deja claro que los principios de protección de datos no serán aplicados a los datos anonimizados debido a que el interesado ya no es identificable. (United Kingdom. Information Commissioner's Office, pág. 11). Esto quiere decir, que una vez que los datos sean propiamente anonimizados, los datos a parte de ya no representar información personal identificable tampoco deberán requerir de principios de protección de datos.

Adicionalmente, es importante comentar en que consiste cada técnica, por lo que, para empezar, se debe reiterar que las técnicas que se considerarán para tratar las variables identificadas como información personal identificable son las siguientes:

- Supresión
- Enmascaramiento

La técnica de supresión consiste en la eliminación de la información personal identificable, es decir, esta técnica consiste en la supresión de todos los datos que contienen las

variables con posible información identificable de los reportes. Esta acción es considerada principalmente cuando “está muy identificada y no puede aplicarse a otro método de protección. También puede quitar una variable cuando es demasiado sensible para uso público o irrelevante a efectos analíticos” (Colombia DANE DIRPEN, 2014, pág. 24). En conclusión, es una técnica enfocada en la preservación de información que está sumamente ligada a la identificación de un estudiante o docente, y, a su vez no presenta relevancia para el enfoque de la investigación.

Por otra parte, la técnica de enmascaramiento, en contraparte con la técnica de supresión, no elimina los datos que contiene esta variable en el reporte, en cambio, se realiza un proceso de enmascaramiento sobre sus datos; en otras palabras, esta técnica modifica los datos originales manteniendo su identidad, pero de tal manera que ya no permitirían la identificación de estudiantes o docentes relacionados a estos datos, por ejemplo, la identificación de estudiante dada por su código podría ser reemplazado por una letra “E” seguido de un número aleatorio. En conclusión, el “enmascaramiento de datos es un método con el que se puede crear una versión que tiene una estructura similar a la de los datos originales pero que no es auténtica” (PowerData, 2023).

Ahora, a partir de la introducción dada a la sección y las conclusiones de la sección anterior, se recapitularán las variables que serán anonimizadas por reporte:

- **Variables por anonimizarse del reporte de asignaturas perdidas:**
 - CEDULA
 - IDD
 - PIDM

- APELLIDO
- SEGUNDO APELLIDO
- NOMBRE
- NRC
- **Variables por anonimizarse del reporte de calificaciones por carrera**
 - pidm_estudiante
 - id_banner_estudiante
 - cedula_pasaporte_estudiante
 - apellidos_estudiante
 - nombres_estudiante
 - NRC
 - pidm_docente
 - id_banner_docente
 - cedula_pasaporte_docente
 - apellidos_docente
 - nombres_docente
 - código_comentario_nota
 - desc_comentario_nota

- **Variables por anonimizarse del reporte de becas**

- PIDM
- ID_BANNER
- APELLIDOS
- PRIMER_NOMBRE
- SEGUNDO_NOMBRE
- CEDULA
- TELEFONO
- CELULAR
- CORREO_PERSONAL
- CORREO_INSTITUC
- CODIGO_COLEGIO

Una vez recapitulado todas las variables que deberán ser sujetas a técnicas de procesamiento para su anonimización, se debe mencionar que no se mostrará el código utilizado para su realización debido a que se corre el riesgo de poner en peligro la información personal de estudiantes y docentes, debido a que, al visualizarse los diferentes algoritmos utilizados se podría efectuar procesos de ingeniería inversa que, a pesar de ser complicado, no sería imposible el recuperar los datos originales.

Sin embargo, de manera general se puede describir las acciones tomadas sobre las variables con respecto a las técnicas utilizadas y consideradas en la sección anterior:

- Con respecto a la técnica de supresión, las variables únicamente fueron eliminadas de los marcos de datos con los que se trabajará el resto de la fase de comprensión de los datos.
- Por otra parte, con respecto a la técnica de enmascaramiento, se realizó un proceso de modificación que concluyo con códigos como “E111” para los estudiantes, y códigos como “D111” para los docentes respectivos.

A partir de la realización de la anonimización, es factible una visualización de los datos de cada reporte, sin embargo, es importante que el lector conozca que la visualización siguiente es superflua pues no tiene el objetivo de mostrar todos los datos, su objetivo consiste en únicamente brindar al lector una vista genérica de los marcos de datos resultantes de la anonimización, es decir, exclusivamente se podrá visualizar un periodo por cada tipo de reporte de datos

Ilustración 23

Visualización de la primera parte del primer periodo del reporte de Asignaturas Perdidas Anonimizado

```
perdidas[0].head()
```

	IDD	PERIODO	PLAN_E	CAMP	DESC_FACULTAD	PROGRAMA	DESC_PROGRAMA	NIVEL_ESTUDIANTE	NF	MATE	CRSE	SHRTCKN_CRSE_TITLE	
0	E5214	202161	2	QUI	Facultad Psicología	Q141	Psicología Clínica		09	65	9	Q0010	Psicoterapia de Adultos
1	E4739	202161	3	QUI	Facultad Cs Humanas	Q076	Sociología Política		04	65	3	Q0921	Proce. Soci. Polít. Ame. Lati.
2	E10373	202161	1	QUI	Facultad Cs Adm. y Contables	Q215	Negocios Internacionales		04	65	4	Q0236	Fundamentals Business Communic
3	E2887	202161	3	QUI	Facultad Economía	Q206	Economía		04	65	3	Q0842	Microeconomía II
4	E2887	202161	3	QUI	Facultad Economía	Q206	Economía		04	65	3	Q0848	Teorías de Desarrollo

Ilustración 24

Visualización de la parte final del primer periodo del reporte de Asignaturas Perdidas Anonimizado

NOTA_FINAL	APROBADO	VECES	MODO	CRED
28	N	1	G	6
11	N	1	G	3
27	N	1	G	3
23	N	1	G	4
28	N	1	G	4

Ilustración 25

Visualización de la primera parte del primer periodo del reporte de Calificaciones por Carrera Anonimizado

```
calificaciones[0].head()
```

	id_banner_estudiante	codigo_periodo	desc_periodo	codigo_facultad	facultad	codigo_carrera	carrera	numero_curso	titulo_curso
0	E1567	202161	Primer P. Ord. 2021 Grado	2	Facultad Cs Adm. y Contables	Q02F	Fac C. Admi Y Contab	Q0521	MATEMÁTICA APLICADA
1	E1567	202161	Primer P. Ord. 2021 Grado	2	Facultad Cs Adm. y Contables	Q02F	Fac C. Admi Y Contab	Q0521	MATEMÁTICA APLICADA
2	E1567	202161	Primer P. Ord. 2021 Grado	2	Facultad Cs Adm. y Contables	Q02F	Fac C. Admi Y Contab	Q0521	MATEMÁTICA APLICADA
3	E1567	202161	Primer P. Ord. 2021 Grado	2	Facultad Cs Adm. y Contables	Q02F	Fac C. Admi Y Contab	Q0521	MATEMÁTICA APLICADA
4	E1567	202161	Primer P. Ord. 2021 Grado	2	Facultad Cs Adm. y Contables	Q02F	Fac C. Admi Y Contab	Q0221	CONTABILIDAD BÁSICA

Ilustración 26

Visualización de la segunda parte del primer periodo del reporte de Calificaciones por Carrera Anonimizado

```
calificaciones[0].iloc[:,9:].head()
```

	numero_paralelo	nombre_parcial	descripcion_parcial	peso	nota_maxima	nota_parcial	porcentaje_nota_parcial	nota_final	id_banner_docente
0	1	PAR. 1	Primer Parcial	25	50	35,93	71,86	35,68	D162
1	1	PAR. 2	Segundo Parcial	25	50	41,08	82,16	35,68	D162
2	1	PAR. 3	Tercer Parcial	25	50	28,38	56,76	35,68	D162
3	1	FINAL	Evaluación Final	25	50	37,33	74,66	35,68	D162
4	1	PAR. 1	Primer Parcial	25	50	34,81	69,62	34,79	D474

Ilustración 27

Visualización de la tercera parte del primer periodo del reporte de Calificaciones por Carrera Anonimizado

```
calificaciones[0].iloc[:,17:].head()
```

	id_banner_docente	codigo_materia	descripcion_materia	secuencial_parcial	fecha_nota_parcial	fecha_nota_final	codigo_programa	descripcion_programa
0	D162	5	Cien.Natur.Matemát.y Estadíst.	1.0	26/3/2021	5/7/2021 14:35	Q179	Administración de Empresas
1	D162	5	Cien.Natur.Matemát.y Estadíst.	2.0	30/4/2021	5/7/2021 14:35	Q179	Administración de Empresas
2	D162	5	Cien.Natur.Matemát.y Estadíst.	3.0	28/5/2021	5/7/2021 14:35	Q179	Administración de Empresas
3	D162	5	Cien.Natur.Matemát.y Estadíst.	4.0	18/6/2021	5/7/2021 14:35	Q179	Administración de Empresas
4	D474	4	Administración	1.0	26/3/2021	5/7/2021 14:35	Q179	Administración de Empresas

Ilustración 28

Visualización de la tercera parte del primer periodo del reporte de Calificaciones por Carrera Anonimizado

```
calificaciones[0].iloc[:,23:].head()
```

	codigo_programa	descripcion_programa	nota_final_SHATCKG	cod_razon_de_cambio	desc_razon_de_cambio	calificacion_del_parcial_score
0	Q179	Administración de Empresas	36	RD	RECALCULO REDONDEO	50
1	Q179	Administración de Empresas	36	RD	RECALCULO REDONDEO	50
2	Q179	Administración de Empresas	36	RD	RECALCULO REDONDEO	50
3	Q179	Administración de Empresas	36	RD	RECALCULO REDONDEO	50
4	Q179	Administración de Empresas	35	RD	RECALCULO REDONDEO	50

Ilustración 29

Visualización de la parte final del primer periodo del reporte de Calificaciones por Carrera Anonimizado

```
calificaciones[0].iloc[:,28:].head()
```

calificacion_del_parcial_score	
0	50
1	50
2	50
3	50
4	50

Ilustración 30

Visualización de la primera parte del primer periodo del reporte de Becas Anonimizado

```
becas[0].head()
```

	ID_BANNER	NACIONALIDAD	GENERO	ETNIA	TIPO_COLEGIO	PROGRAMA	PROGAMA_DESC	CAMPUS_PROGRAMA	AREA_MATERIAS
0	E5128	Ecuatoriano	Mujer	NaN	NaN	Q148	Enfermería	QUI	Q148-01
1	E11319	Ecuatoriano	Hombre	Mestizo	NaN	Q206	Economía	QUI	Q206-07
2	E10709	Ecuatoriano	Mujer	NaN	Part L	Q215	Negocios Internacionales	QUI	Q215-04
3	E7918	Ecuatoriano	Mujer	Mestizo	NaN	Q001	Arquitectura	QUI	Q001-10
4	E10503	Ecuatoriano	Hombre	NaN	NaN	Q193	Teología	QUI	Q193-02

Ilustración 31

Visualización de la segunda parte del primer periodo del reporte de Becas Anonimizado

```
becas[0].head()
```

NIVEL_POR_MATERIAS	NIVEL_POR_ASIG_APROBADAS	CATEGORIA_BECA_SOCIOECONOMICA	TIPO_REGIMEN	NUM_MATERIAS_SEGUNDA_MATRICULA
01	2.0	A	NaN	1
07	8.0	C	NaN	0
04	7.0	C	NaN	1
10	10.0	A	NaN	0
02	2.0	B	NaN	0

Ilustración 32

Visualización de la parte final del primer periodo del reporte de Becas Anonimizado

NUM_MATERIAS_TERCERA_MATRICULA	TIENE_PAGO	TIENE_SALDO
0	SI	NO
0	SI	NO
0	SI	NO
0	SI	NO
0	SI	NO

En conclusión, a pesar de que no haya sido factible la demostración de los algoritmos utilizados en la anonimización de las variables de cada reporte, en la visualización del primer periodo de cada tipo de reporte se evidencia que los procesos de anonimización fueron exitosos.

4.1.8. Identificación de tipos de variables

En esta sección se realizará la identificación de tipos de variables, es decir, se catalogarán las variables de cada reporte en variables categóricas o numéricas, sin embargo, es necesario realizar una breve comprensión de que se define como variables categóricas y numéricas.

Para empezar, “Si los valores de una variable son solo etiquetas o entre ellos solo se puede establecer un orden que refleja diferentes grados del atributo que se mide; entonces, estos corresponden, respectivamente, a datos cualitativos o categóricos.” (Véliz, 2020, pág. 7), en otras palabras, las variables categóricas se refieren a datos que determinan calidades o presentan un orden o jerarquía de medición de una determinada variable.

En contraste, “Si las diferencias o los cocientes entre los valores de una variable tienen sentido en la realidad, los datos correspondientes son cuantitativos o numéricos, respectivamente” (Véliz, 2020, pág. 7), en otras palabras, las variables numéricas se refieren a datos numéricos con los que se pueden ejecutar operaciones aritméticas. Adicionalmente, las variables numéricas pueden considerarse discretas o continuas dada la siguiente definición: “Un conjunto de datos es discreto si puede enumerarse. Si cualquiera de los valores de un intervalo puede ser valores de una variable, entonces, estos corresponden a datos continuos.” (Véliz, 2020, pág. 8).

Además, es crucial entender que el proceso de identificación de los respectivos tipos de variables, ya sean categóricas o continuas, constituye la piedra angular para realizar cualquier tipo de exploración de datos. Esto no solo incluye la comprensión de las distribuciones individuales de las variables, sino también la capacidad de examinar las interacciones y las relaciones entre ellas.

Sin embargo, para lograr la correcta identificación de los tipos de variables se presenció la necesidad de evaluar el contenido de cada una de las variables que posee cada reporte, sobre todo después de haberse efectuado un proceso de anonimización con la utilización de técnicas como la de supresión. Por lo que, primero se realizó una recapitulación de las variables existentes en cada reporte:

Ilustración 33

Recapitulación de variables del reporte de Asignaturas Perdidas

```
print(f"El reporte de asignaturas perdidas posee las siguientes variables: ")
for column in perdidas[0].columns: print("\t- ", column)
```

El reporte de asignaturas perdidas posee las siguientes variables:

- IDD
- PERIODO
- PLAN_E
- CAMP
- DESC_FACULTAD
- PROGRAMA
- DESC_PROGRAMA
- NIVEL_ESTUDIANTE
- NF
- MATE
- CRSE
- SHRTCKN_CRSE_TITLE
- NOTA_FINAL
- APROBADO
- VECES
- MODO
- CRED

Ilustración 34

Recapitulación de las variables del reporte de Calificaciones por Carrera

```
print(f"El reporte de asignaturas perdidas posee las siguientes variables: ")
for column in calificaciones[0].columns: print("\t- ", column)
```

El reporte de asignaturas perdidas posee las siguientes variables:

- id_banner_estudiante
- codigo_periodo
- desc_periodo
- codigo_facultad
- facultad
- codigo_carrera
- carrera
- numero_curso
- titulo_curso
- numero_paralelo
- nombre_parcial
- descripcion_parcial
- peso
- nota_maxima
- nota_parcial
- porcentaje_nota_parcial
- nota_final
- id_banner_docente
- codigo_materia
- descripcion_materia
- secuencial_parcial
- fecha_nota_parcial
- fecha_nota_final
- codigo_programa
- descripcion_programa
- nota_final_SHATCKG
- cod_razon_de_cambio
- desc_razon_de_cambio
- calificacion_del_parcial_score

Ilustración 35

Recapitulación de las variables del reporte de Becas

```
print(f"El reporte de asignaturas perdidas posee las siguientes variables: ")
for column in becas[0].columns: print("\t- ", column)
```

El reporte de asignaturas perdidas posee las siguientes variables:

- ID_BANNER
- NACIONALIDAD
- GENERO
- ETNIA
- TIPO_COLEGIO
- PROGRAMA
- PROGRAM_DESC
- CAMPUS_PROGRAMA
- AREA_MATERIAS
- NIVEL_POR_MATERIAS
- NIVEL_POR_ASIG_APROBADAS
- CATEGORIA_BECA_SOCIOECONOMICA
- TIPO_REGIMEN
- NUM_MATERIAS_SEGUNDA_MATRICULA
- NUM_MATERIAS_TERCERA_MATRICULA
- TIENE_PAGO
- TIENE_SALDO

Ahora, una vez recordado las variables disponibles por cada respectivo tipo de reporte, se realizó un análisis de los contenidos únicos que ostentan cada una de las variables debido a que el conocer el significado o sentido de una variable podría no ser suficiente para definir su tipo, es decir, se realizó un código que permitió el análisis de los datos no repetidos que poseen cada una de las variables recapituladas. Los códigos pertinentes se podrán visualizar en las siguientes ilustraciones, sin embargo, los resultados de estos códigos únicamente se mostrarán como conclusiones, debido a que, los códigos implementados dieron como resultado un despliegue de una cantidad inmensa de datos, por lo que, resulta contraproducente el desplegar una avalancha de ilustraciones mostrando partes de los resultados; no obstante, si el distinguido lector tiene curiosidad de los resultados de los códigos, es cordialmente invitado a visualizar el anexo “Resultado de los códigos para la identificación de variables únicas en los reportes.docx”, en

este, se podrá visualizar a detalle los resultados que se presentaron a partir de los siguiente códigos:

Ilustración 36

Código utilizado para el análisis del contenido de las variables del reporte de Asignaturas Perdidas

7.1.3. Análisis avanzado de los contenidos de cada variable

```
'''
Despliegue de información de cada característica de cada periodo
'''

for df in perdidas:
    print("\n\n" + "*" * 100, end="\n\n")
    print("Analizando el dataframe: ", df.name, end="\n\n")
    for column in df.columns:
        if column != "IDD":
            print(f"Analizando la característica '{column}':\n")
            print(f"Cantidad de registros únicos: {len(df[column].unique())}")
            print(f"Registros únicos de la característica {column}:\n\t", df[column].unique(), end="\n\n")
```

Ilustración 37

Código utilizado para el análisis del contenido de las variables del reporte de Calificaciones por Carrera

7.2.3. Análisis avanzado de los contenidos de cada variable

```
'''
Despliegue de información de cada característica de cada periodo
'''

for df in calificaciones:
    print("\n\n" + "*" * 100, end="\n\n")
    print("Analizando el dataframe: ", df.name, end="\n\n")
    for column in df.columns:
        if column != "id_banner_estudiante":
            print(f"Analizando la característica '{column}':\n")
            print(f"Cantidad de registros únicos: {len(df[column].unique())}")
            print(f"Registros únicos de la característica {column}:\n\t", df[column].unique(), end="\n\n")
```

Ilustración 38

Código utilizado para el análisis del contenido de las variables del reporte de Becas

7.3.3. Análisis avanzado de los contenidos de cada variable

```
...
Despliegue de información de cada característica de cada periodo
...

for df in becas:
    print("\n\n" + "*" * 100, end="\n\n")
    print("Analizando el dataframe: ", df.name, end="\n\n")
    for column in df.columns:
        if column != "ID_BANNER":
            print(f"Analizando la característica '{column}':\n")
            print(f"Cantidad de registros únicos: {len(df[column].unique())}")
            print(f"Registros únicos de la característica {column}:\n\t", df[column].unique(), end="\n\n")
```

Una vez se ejecutaron los respectivos códigos para el análisis de los datos de cada variable, se logró identificar apropiadamente los tipos de datos que cada variable se jacta, los mismo, se podrán visualizar con su respectiva razón en las siguientes tablas, comenzando con la exhibición de la clasificación de las variables del reporte de asignaturas perdidas:

Tabla 9

Clasificación de Variables del Reporte de Asignaturas Perdidas

Característica	Tipo	Subtipo	Razón
IDD	categorica	nominal	Identificador único del estudiante en el banner, no existe un orden o jerarquía inherente.
PERIODO	categorica	ordinal	Representa un periodo académico y puede ser ordenado cronológicamente.
PLAN_E	categorica	nominal	Código del plan de estudios al que pertenece el estudiante, no hay un orden específico.
CAMP	categorica	nominal	Campus al que pertenece el estudiante, no hay un orden o jerarquía inherente.

DESC_FACULTAD	categorica	nominal	Nombre completo de la facultad a la que pertenece el estudiante, no hay un orden específico.
PROGRAMA	categorica	nominal	Código del programa académico al que pertenece el estudiante, no hay un orden específico.
DESC_PROGRAMA	categorica	nominal	Nombre del programa en el que se encuentra inscrito el estudiante, no hay un orden específico.
NIVEL_ESTUDIANTE	categorica	ordinal	Nivel académico del estudiante dentro del programa, se puede ordenar según la progresión educativa.
NF	categorica	nominal	Nivel académico en el que se encuentra el estudiante, puede ser Grado, Idiomas, Técnico Superior, etc.
MATE	categorica	nominal	Código del área a la que pertenece la asignatura, no hay un orden o jerarquía inherente.
CRSE	categorica	nominal	Número de la asignatura, no hay un orden o jerarquía inherente.
SHRTCKN_CRSE_TITLE	categorica	nominal	Nombre de la asignatura perdida, no hay un orden específico.
NOTA_FINAL	numérica	continua	Calificación final del estudiante en el curso, es una medida continua.
APROBADO	categorica	nominal	Indicador si el estudiante aprobó o reprobó el curso, no hay un orden o jerarquía inherente.

VECES	numérica	discreta	Número de veces que el estudiante ha tomado el curso, es una medida discreta.
MODO	categórica	nominal	Escala de calificación, puede ser G para General, no hay un orden o jerarquía inherente.
CRED	numérica	discreta	Número de créditos asignados al curso, es una medida discreta.

En segundo lugar, se procederá a demostrar la tabla con la clasificación de las variables del reporte de calificaciones por carrera:

Tabla 10
Clasificación de Variables del Reporte de Calificaciones por Carrera

Característica	Tipo	Subtipo	Razón
id_banner_estudiante	categórica	nominal	Identificador único del estudiante en el sistema Banner, no existe un orden o jerarquía inherente.
codigo_periodo	categórica	ordinal	Representa un periodo académico y puede ser ordenado cronológicamente.
desc_periodo	categórica	ordinal	Representa el nombre del periodo académico y puede ser ordenado cronológicamente.
facultad	categórica	nominal	Nombre de la facultad a la que pertenece el estudiante, no hay un orden específico.

codigo_carrera	categorica	nominal	Código de la carrera a la que pertenece el estudiante, no hay un orden específico.
carrera	categorica	nominal	Nombre de la carrera a la que pertenece el estudiante, no hay un orden específico.
numero_curso	categorica	nominal	Número del curso específico que el estudiante tomó, no hay un orden o jerarquía inherente.
titulo_curso	categorica	nominal	Nombre del curso específico que el estudiante tomó, no hay un orden específico.
numero_paralelo	categorica	nominal	Número de paralelo en el que el estudiante tomó el curso, no hay un orden o jerarquía inherente.
nombre_parcial	categorica	nominal	Nombre del parcial del que se tomó el registro de la calificación del estudiante, no hay un orden específico.
descripcion_parcial	categorica	nominal	Descripción del parcial del que se tomó el registro de la calificación del estudiante, no hay un orden específico.
peso	categorica	ordinal	Representa el peso que cada parcial posee con respecto a la nota final.
nota_maxima	constante		Representa la nota máxima que las calificaciones pueden alcanzar.
nota_parcial	numérica	continua	Calificación que el estudiante obtuvo en el parcial, es una medida continua.

porcentaje_nota_parcial	numérica	continua	Porcentaje de la nota parcial con respecto a la nota máxima, es una medida continua.
nota_final	numérica	continua	Calificación final del estudiante en el curso, es una medida continua.
id_banner_docente	categórica	nominal	Identificador único del docente asociado al registro, no existe un orden o jerarquía inherente.
codigo_materia	categórica	nominal	Código de la materia cursada, no hay un orden o jerarquía inherente.
descripcion_materia	categórica	nominal	Área de la materia cursada, no hay un orden o jerarquía inherente.
secuencial_parcial	Categórica	Ordinal	Número secuencial asignado a la evaluación específica (por ejemplo, primer parcial (1), segundo parcial (2), etc.).
fecha_nota_parcial	Categórica	Ordinal	Fecha en la que se registró la calificación de la evaluación específica.
fecha_nota_final	Categórica	Ordinal	Fecha y hora en la que se registró la nota final del curso.
codigo_programa	Categórica	Nominal	Código del programa en el que se encuentra inscrito el estudiante, no hay un orden específico.
descripcion_programa	Categórica	Nominal	Nombre del programa en el que se encuentra inscrito el estudiante, no hay un orden específico.
nota_final_SHATCKG	Categórica	Ordinal	Nota final según el registro oficial de calificaciones, sin embargo, en sus

datos se encontraron registros con “AP”, lo que sugiere que estos datos no corresponden a únicamente valores numéricos y podrían referirse a estados de aprobación.

cod_razon_de_cambio	Categórica	Nominal	Código de la razón del cambio de la calificación, no hay un orden o jerarquía inherente.
desc_razon_de_cambio	Categórica	Nominal	Descripción de la razón del cambio de la calificación, no hay un orden o jerarquía inherente.
calificacion_del_parcial_score	Constante		Calificación máxima de una evaluación específica, puede ser constante ya que representa la máxima calificación posible.

Finalmente, se podrá visualizar la clasificación de las variables del reporte de becas a continuación:

Tabla 11
Clasificación de Variables del Reporte de Becas

Característica	Tipo	Sub-tipo	Razón
ID_BANNER	categórica	nominal	Identificador único del estudiante en el sistema Banner, no existe un orden o jerarquía inherente.

NACIONALIDAD	categoría	nominal	Nacionalidad del estudiante, no hay un orden o jerarquía inherente.
GENERO	categoría	nominal	Género del estudiante, no hay un orden o jerarquía inherente.
ETNIA	categoría	nominal	Etnia a la que pertenece el estudiante, no hay un orden o jerarquía inherente.
TIPO_COLEGIO	categoría	nominal	Tipo de colegio donde el estudiante completó su educación secundaria, no hay un orden o jerarquía inherente.
PROGRAMA	categoría	nominal	Código de la carrera a la que pertenece el estudiante, no hay un orden específico.
PROGRAMA_DESC	categoría	nominal	Nombre de la carrera a la que pertenece el estudiante, no hay un orden específico.

CAMPUS_PROGRAMA	categoría	nominal	Campus en el que se encuentra el estudiante, no hay un orden o jerarquía inherente.
AREAS_MATERIAS	categoría	nominal	Identificador del área y curso durante ese periodo, no hay un orden o jerarquía inherente.
NIVEL_POR_MATERIAS	categoría	ordinal	Nivel del estudiante en relación con las materias que está cursando, se puede ordenar según la progresión académica.
NIVEL_POR_ASIG_APROBADAS	categoría	ordinal	Nivel de la asignatura aprobada por el estudiante, se puede ordenar según la progresión académica.
CATEGORIA_BECA_SOCIOECONOMICA	categoría	ordinal	Categoría de la beca socioeconómica del estudiante, se puede ordenar según el nivel de apoyo económico otorgado.

TIPO_REGIMEN	categoría	nominal	Tipo de régimen de beca asignado al estudiante, no hay un orden o jerarquía inherente.
NUM_MATERIAS_SEGUNDA_MATRICULA	numérica	discreta	Número de materias en la segunda matrícula del estudiante, es una medida discreta.
NUM_MATERIAS_TERCERA_MATRICULA	numérica	discreta	Número de materias en la tercera matrícula del estudiante, es una medida discreta.
TIENE_PAGO	categoría	nominal	Indica si el estudiante tiene un pago registrado, no hay un orden o jerarquía inherente.
TIENE_SALDO	categoría	nominal	Indica si el estudiante tiene un saldo pendiente, no hay un orden o jerarquía inherente.

En conclusión, del análisis realizado se determinó el tipo de variables que deberían poseer cada una de las disponibles en sus asignados reportes, sin embargo, el trabajo en esta sección está lejos de terminar, pues, es necesario realizar un ajuste de los tipos de datos de

aquellas variables que, dado el análisis, no se corresponden a sus tipos de variables actuales. Para comenzar, se realizó un código capaz de permitir una comparativa entre los tipos de datos que hasta este momento poseen las variables con respecto a los tipos identificados, el código contiene tres funciones distintas que permitirán la visualización de las comparativas al momento de su llamado. Los códigos para la generación de las respectivas comparativas se podrán visualizar a continuación:

Ilustración 39

Código para la comparativa de tipos de variables del reporte de Asignaturas Perdidas

```
def comparativa_tipos_perdidas(perdidas_dataframes):  
    ...  
    Comparativa de tipos de datos actuales con los que se definieron en el análisis  
    ...  
    # Tipos de datos que se definieron en el análisis  
    dict_tipos = {  
        "IDD" : object,  
        "PERIODO" : object,  
        "PLAN_E" : object,  
        "CAMP" : object,  
        "DESC_FACULTAD" : object,  
        "PROGRAMA" : object,  
        "DESC_PROGRAMA" : object,  
        "NIVEL_ESTUDIANTE" : object,  
        "NF" : object,  
        "MATE" : object,  
        "CRSE" : object,  
        "SHRTCKN_CRSE_TITLE" : object,  
        "NOTA_FINAL" : float,  
        "APROBADO" : object,  
        "VECES" : int,  
        "MOD0" : object,  
        "CRED" : float  
    }  
    # Transformación de dict a Series de Los tipos  
    serie_tipos = pd.Series(dict_tipos, name="tipos")  
  
    # Despliegue de información general  
    for df in perdidas_dataframes:  
        print("\n\n" + "*" * 100, end="\n\n")  
        print("Analizando el dataframe: ", df.name, end="\n\n")  
        print(df.dtypes == serie_tipos)  
        print("\nCaracterísticas que no poseen los tipos de datos requeridos: ", end="")  
        print([df.columns[x] for x in range(len(df.dtypes)) if df.dtypes[x] != serie_tipos[x]])
```

Ilustración 40

Código para la comparativo de los tipos de variables del reporte de Calificaciones por Carrera

```
def comparativa_tipos_calificaciones(calificaciones_dataframes):
    """
    Comparativa de tipos de datos actuales con los que se definieron en el análisis
    """
    # Tipos de datos que se definieron en el análisis
    dict_tipos = {
        "id_banner_estudiante" : object,
        "codigo_periodes" : object,
        "desc_periodes" : object,
        "codigo_facultad" : object,
        "facultad" : object,
        "codigo_carrera" : object,
        "carrera" : object,
        "numero_curso" : object,
        "titulo_curso" : object,
        "numero_paralelo" : object,
        "nombre_parcial" : object,
        "descripcion_parcial" : object,
        "peso" : object,
        "nota_maxima" : object,
        "nota_parcial" : float,
        "porcentaje_nota_parcial" : float,
        "nota_final" : float,
        "id_banner_docente" : object,
        "codigo_materia" : object,
        "descripcion_materia" : object,
        "secuencial_parcial" : object,
        "fecha_nota_parcial" : object,
        "fecha_nota_final" : object,
        "codigo_programa" : object,
        "descripcion_programa" : object,
        "nota_final_SHATCKG" : object,
        "cod_razon_de_cambio" : object,
        "desc_razon_de_cambio" : object,
        "calificacion_del_parcial_score" : object,
    }
    # Transformación de dict a Series de los tipos
    serie_tipos = pd.Series(dict_tipos, name="tipos")

    # Despliegue de información general
    for df in calificaciones_dataframes:
        print("\n\n" + "*" * 100, end="\n\n")
        print("Analizando el dataframe: ", df.name, end="\n\n")
        print(df.dtypes == serie_tipos)
        print("\nCaracterísticas que no poseen los tipos de datos requeridos: ", end="")
        print([df.columns[x] for x in range(len(df.dtypes)) if df.dtypes[x] != serie_tipos[x]])
```


Ilustración 41

Código para la comparativa de los tipos de variables del reporte de Becas

```
def comparativa_tipos_becas(becas_dataframes):  
    ...  
    Comparativa de tipos de datos actuales con los que se definieron en el análisis  
    ...  
    # Tipos de datos que se definieron en el análisis  
    dict_tipos = {  
        "ID_BANNER" : object,  
        "NACIONALIDAD" : object,  
        "GENERO" : object,  
        "ETNIA" : object,  
        "TIPO_COLEGIO" : object,  
        "PROGRAMA" : object,  
        "PROGAMA_DESC" : object,  
        "CAMPUS_PROGRAMA" : object,  
        "AREA_MATERIAS" : object,  
        "NIVEL_POR_MATERIAS" : object,  
        "NIVEL_POR_ASIG_APROBADAS" : object,  
        "CATEGORIA_BECA_SOCIOECONOMICA" : object,  
        "TIPO_REGIMEN" : object,  
        "NUM_MATERIAS_SEGUNDA_MATRICULA" : int,  
        "NUM_MATERIAS_TERCERA_MATRICULA" : int,  
        "TIENE_PAGO" : object,  
        "TIENE_SALDO" : object  
    }  
    # Transformación de dict a Series de los tipos  
    serie_tipos = pd.Series(dict_tipos, name="tipos")  
  
    # Despliegue de información general  
    for df in becas_dataframes:  
        print("\n\n" + "*" * 100, end="\n\n")  
        print("Analizando el dataframe: ", df.name, end="\n\n")  
        print(df.dtypes == serie_tipos)  
        print("\nCaracterísticas que no poseen los tipos de datos requeridos: ", end="")  
        print([df.columns[x] for x in range(len(df.dtypes)) if df.dtypes[x] != serie_tipos[x]])
```

A partir de los códigos presentados, antes de presentar sus resultados, es necesario remarcar lo que se espera obtener, dicho esto, en las ilustraciones consiguientes se podrán observar los nombres de las diferentes variables al lado de valores como “True” y “False”, estos valores indican si el tipo de dato de la variable corresponde al tipo de dato requerido y establecido durante la identificación de los tipos de datos, es decir, sabiendo que el código realiza una comparación entre el tipo de dato actual de las variables con respecto al tipo de dato idóneo, los resultados posibles son “True” para simbolizar que ambas evaluaciones se

corresponden, mientras que, “False” es empleado para demostrar que existe un problema con el tipo de dato actual de la variable y que necesita de un ajuste.

Ilustración 42

Ejecución de la comparativa del reporte de Asignaturas Perdidas

```
comparativa_tipos_perdidas(perdidas)
```

Ilustración 43

Resultados de la comparativa del primer periodo del reporte de Asignaturas Perdidas

```
*****
```

```
Analizando el dataframe: perdidas01
```

IDD	True
PERIODO	False
PLAN_E	False
CAMP	True
DESC_FACULTAD	True
PROGRAMA	True
DESC_PROGRAMA	True
NIVEL_ESTUDIANTE	True
NF	False
MATE	False
CRSE	True
SHRTCKN_CRSE_TITLE	True
NOTA_FINAL	False
APROBADO	True
VECES	False
MODO	True
CRED	False

```
dtype: bool
```

```
Características que no poseen los tipos de datos requeridos: ['PERIODO', 'PLAN_E', 'NF', 'MATE', 'NOTA_FINAL', 'VECES', 'CRED']
```

Ilustración 44

Resultados de la comparativa del segundo periodo del reporte de Asignaturas Perdidas

```
*****

Analizando el dataframe: perdidas02

IDD                True
PERIODO            False
PLAN_E             False
CAMP               True
DESC_FACULTAD      True
PROGRAMA           True
DESC_PROGRAMA      True
NIVEL_ESTUDIANTE   True
NF                 False
MATE               False
CRSE               True
SHRTCKN_CRSE_TITLE True
NOTA_FINAL         False
APROBADO           True
VECES              False
MOD0               True
CRED               False
dtype: bool

Características que no poseen los tipos de datos requeridos: ['PERIODO', 'PLAN_E', 'NF', 'MATE', 'NOTA_FINAL', 'VECES', 'CRED']
```

Ilustración 45

Resultados de la comparativo del tercer periodo del reporte de Asignaturas Perdidas

```
Analizando el dataframe: perdidas03

IDD                True
PERIODO            False
PLAN_E             False
CAMP               True
DESC_FACULTAD      True
PROGRAMA           True
DESC_PROGRAMA      True
NIVEL_ESTUDIANTE   True
NF                 False
MATE               False
CRSE               True
SHRTCKN_CRSE_TITLE True
NOTA_FINAL         False
APROBADO           True
VECES              False
MOD0               True
CRED               False
dtype: bool

Características que no poseen los tipos de datos requeridos: ['PERIODO', 'PLAN_E', 'NF', 'MATE', 'NOTA_FINAL', 'VECES', 'CRED']

*****
```

Ilustración 46

Resultados de la comparativa del cuarto periodo del reporte de Asignaturas Perdidas

```
*****  
Analizando el dataframe: perdidas04  
  
IDD                True  
PERIODO            False  
PLAN_E             False  
CAMP               True  
DESC_FACULTAD      True  
PROGRAMA           True  
DESC_PROGRAMA      True  
NIVEL_ESTUDIANTE   False  
NF                 False  
MATE               False  
CRSE               True  
SHRTCKN_CRSE_TITLE True  
NOTA_FINAL         False  
APROBADO           True  
VECES              False  
MOD0               True  
CRED               False  
dtype: bool  
  
Características que no poseen los tipos de datos requeridos: ['PERIODO', 'PLAN_E', 'NIVEL_ESTUDIANTE', 'NF', 'MATE', 'NOTA_FINAL', 'VECES', 'CRED']
```

Ilustración 47

Ejecución de la comparativa del reporte de Calificaciones por Carrera

```
comparativa_tipos_calificaciones(calificaciones)
```

Ilustración 48

Resultados de la comparativa del primer periodo del reporte de Calificaciones por Carrera

Analizando el dataframe: calificaciones01

id_banner_estudiante	True
codigo_periodo	False
desc_periodo	True
codigo_facultad	False
facultad	True
codigo_carrera	True
carrera	True
numero_curso	True
titulo_curso	True
numero_paralelo	True
nombre_parcial	True
descripcion_parcial	True
peso	True
nota_maxima	False
nota_parcial	False
porcentaje_nota_parcial	False
nota_final	False
id_banner_docente	True
codigo_materia	False
descripcion_materia	True
secuencial_parcial	False
fecha_nota_parcial	True
fecha_nota_final	True
codigo_programa	True
descripcion_programa	True
nota_final_SHATCKG	True
cod_razon_de_cambio	True
desc_razon_de_cambio	True
calificacion_del_parcial_score	False

dtype: bool

Características que no poseen los tipos de datos requeridos: ['codigo_periodo', 'codigo_facultad', 'nota_maxima', 'nota_parcial', 'porcentaje_nota_parcial', 'nota_final', 'codigo_materia', 'secuencial_parcial', 'calificacion_del_parcial_score']

Ilustración 49

Resultados de la comparativa del segundo periodo del reporte de Calificaciones por Carrera

Analizando el dataframe: calificaciones02

id_banner_estudiante	True
codigo_periodo	False
desc_periodo	True
codigo_facultad	False
facultad	True
codigo_carrera	True
carrera	True
numero_curso	True
titulo_curso	True
numero_paralelo	True
nombre_parcial	True
descripcion_parcial	True
peso	True
nota_maxima	False
nota_parcial	False
porcentaje_nota_parcial	False
nota_final	False
id_banner_docente	True
codigo_materia	False
descripcion_materia	True
secuencial_parcial	False
fecha_nota_parcial	True
fecha_nota_final	True
codigo_programa	True
descripcion_programa	True
nota_final_SHATCKG	True
cod_razon_de_cambio	True
desc_razon_de_cambio	True
calificacion_del_parcial_score	False

dtype: bool

Características que no poseen los tipos de datos requeridos: ['codigo_periodo', 'codigo_facultad', 'nota_maxima', 'nota_parcial', 'porcentaje_nota_parcial', 'nota_final', 'codigo_materia', 'secuencial_parcial', 'calificacion_del_parcial_score']

Ilustración 50

Resultados de la comparativa del tercer periodo del reporte de Calificaciones por Carrera

Analizando el dataframe: calificaciones03

id_banner_estudiante	True
codigo_periodo	False
desc_periodo	True
codigo_facultad	False
facultad	True
codigo_carrera	True
carrera	True
numero_curso	True
titulo_curso	True
numero_paralelo	True
nombre_parcial	True
descripcion_parcial	True
peso	True
nota_maxima	False
nota_parcial	False
porcentaje_nota_parcial	False
nota_final	False
id_banner_docente	True
codigo_materia	False
descripcion_materia	True
secuencial_parcial	False
fecha_nota_parcial	True
fecha_nota_final	True
codigo_programa	True
descripcion_programa	True
nota_final_SHATCKG	True
cod_razon_de_cambio	True
desc_razon_de_cambio	True
calificacion_del_parcial_score	False

dtype: bool

Características que no poseen los tipos de datos requeridos: ['codigo_periodo', 'codigo_facultad', 'nota_maxima', 'nota_parcial', 'porcentaje_nota_parcial', 'nota_final', 'codigo_materia', 'secuencial_parcial', 'calificacion_del_parcial_score']

Ilustración 51

Resultados de la comparativa del cuarto periodo del reporte de Calificaciones por Carrera

```
Analizando el dataframe: calificaciones04
```

id_banner_estudiante	True
codigo_periodes	False
desc_periodes	True
codigo_facultad	False
facultad	True
codigo_carrera	True
carrera	True
numero_curso	True
titulo_curso	True
numero_paralelo	True
nombre_parcial	True
descripcion_parcial	True
peso	True
nota_maxima	False
nota_parcial	False
porcentaje_nota_parcial	False
nota_final	False
id_banner_docente	True
codigo_materia	False
descripcion_materia	True
secuencial_parcial	False
fecha_nota_parcial	True
fecha_nota_final	True
codigo_programa	True
descripcion_programa	True
nota_final_SHATCKG	True
cod_razon_de_cambio	True
desc_razon_de_cambio	True
calificacion_del_parcial_score	False

dtype: bool

Características que no poseen los tipos de datos requeridos: ['codigo_periodes', 'codigo_facultad', 'nota_maxima', 'nota_parcial', 'porcentaje_nota_parcial', 'nota_final', 'codigo_materia', 'secuencial_parcial', 'calificacion_del_parcial_score']

Ilustración 52

Ejecución de la comparativa del reporte de Becas

```
comparativa_tipos_becas(becas)
```


Ilustración 53

Resultados de la comparativa del primer periodo del reporte de Becas

```
*****

Analizando el dataframe: becas01

ID_BANNER                True
NACIONALIDAD              True
GENERO                    True
ETNIA                     True
TIPO_COLEGIO              True
PROGRAMA                  True
PROGAMA_DESC              True
CAMPUS_PROGRAMA           True
AREA_MATERIAS             True
NIVEL_POR_MATERIAS        True
NIVEL_POR_ASIG_APROBADAS  False
CATEGORIA_BECA_SOCIOECONOMICA  True
TIPO_REGIMEN              False
NUM_MATERIAS_SEGUNDA_MATRICULA  False
NUM_MATERIAS_TERCERA_MATRICULA  False
TIENE_PAGO                True
TIENE_SALDO               True
dtype: bool

Características que no poseen los tipos de datos requeridos: ['NIVEL_POR_ASIG_APROBADAS', 'TIPO_REGIMEN', 'NUM_MATERIAS_SEGUNDA_MATRICULA', 'NUM_MATERIAS_TERCERA_MATRICULA']
```

Ilustración 54

Resultados de la comparativa del segundo periodo del reporte de Becas

```
*****

Analizando el dataframe: becas02

ID_BANNER                True
NACIONALIDAD              True
GENERO                    True
ETNIA                     True
TIPO_COLEGIO              True
PROGRAMA                  True
PROGAMA_DESC              True
CAMPUS_PROGRAMA           True
AREA_MATERIAS             True
NIVEL_POR_MATERIAS        True
NIVEL_POR_ASIG_APROBADAS  False
CATEGORIA_BECA_SOCIOECONOMICA  False
TIPO_REGIMEN              False
NUM_MATERIAS_SEGUNDA_MATRICULA  False
NUM_MATERIAS_TERCERA_MATRICULA  False
TIENE_PAGO                True
TIENE_SALDO               True
dtype: bool

Características que no poseen los tipos de datos requeridos: ['NIVEL_POR_ASIG_APROBADAS', 'CATEGORIA_BECA_SOCIOECONOMICA', 'TIPO_REGIMEN', 'NUM_MATERIAS_SEGUNDA_MATRICULA', 'NUM_MATERIAS_TERCERA_MATRICULA']
```

Ilustración 55

Resultados de la comparativa del tercer periodo del reporte de Becas

```
*****

Analizando el dataframe: becas03

ID_BANNER                True
NACIONALIDAD              True
GENERO                    True
ETNIA                     True
TIPO_COLEGIO              True
PROGRAMA                  True
PROGAMA_DESC              True
CAMPUS_PROGRAMA           True
AREA_MATERIAS              True
NIVEL_POR_MATERIAS         True
NIVEL_POR_ASIG_APROBADAS   False
CATEGORIA_BECA_SOCIOECONOMICA True
TIPO_REGIMEN              True
NUM_MATERIAS_SEGUNDA_MATRICULA False
NUM_MATERIAS_TERCERA_MATRICULA False
TIENE_PAGO                 True
TIENE_SALDO                True
dtype: bool

Características que no poseen los tipos de datos requeridos: ['NIVEL_POR_ASIG_APROBADAS', 'NUM_MATERIAS_SEGUNDA_MATRICULA', 'NUM_MATERIAS_TERCERA_MATRICULA']
```

Ilustración 56

Resultados de la comparativa del cuarto periodo del reporte de Becas

```
*****

Analizando el dataframe: becas04

ID_BANNER                True
NACIONALIDAD              True
GENERO                    True
ETNIA                     True
TIPO_COLEGIO              True
PROGRAMA                  True
PROGAMA_DESC              True
CAMPUS_PROGRAMA           True
AREA_MATERIAS              True
NIVEL_POR_MATERIAS         True
NIVEL_POR_ASIG_APROBADAS   False
CATEGORIA_BECA_SOCIOECONOMICA False
TIPO_REGIMEN              False
NUM_MATERIAS_SEGUNDA_MATRICULA False
NUM_MATERIAS_TERCERA_MATRICULA False
TIENE_PAGO                 True
TIENE_SALDO                True
dtype: bool

Características que no poseen los tipos de datos requeridos: ['NIVEL_POR_ASIG_APROBADAS', 'CATEGORIA_BECA_SOCIOECONOMICA', 'TIPO_REGIMEN', 'NUM_MATERIAS_SEGUNDA_MATRICULA', 'NUM_MATERIAS_TERCERA_MATRICULA']
```

En conclusión, las comparativas realizadas demostraron la existencia de inconsistencias con lo que respecta a los tipos de datos que actualmente poseen las variables con el especificado

en el análisis inicial, esto implica que pueden existir posibles errores durante el proceso de exploración de variables numéricas y categóricas, debido a que “es importante identificar el tipo de datos, para saber qué tipo de tratamiento se le dará (...) y que tipo de gráfico se utilizará.” (Loayza & Loayza, s. f.) Es decir, es de vital importancia identificar correctamente los tipos de datos de las variables para poder emplear gráficos adecuados sobre ellas, lo que significa que esta representación debe adecuarse igualmente a como se encuentran definidas estas variables en el código, puesto que los gráficos empleados en variables numéricas requieren que sus datos sean numéricos. Por lo que, se ve la necesidad de realizar un proceso de ajuste de las variables disponibles en cada reporte, en específico, se debe realizar un ajuste a las siguientes variables:

- **Variables inconsistentes en el reporte de asignaturas perdidas:**

- PERIODO
- PLAN_E
- NIVEL_ESTUDIANTE
- NF
- MATE
- NOTA_FINAL
- VECES
- CRED

- **Variables inconsistentes en el reporte de asignaturas perdidas:**

- codigo_periodo

- codigo_facultad
- nota_maxima
- nota_parcial
- porcentaje_nota_parcial
- nota_final
- codigo_materia
- secuencial_parcial
- nota_final_SHATCKG
- calificacion_del_parcial_score
- **Variables inconsistentes en el reporte de becas**
 - NIVEL_POR_ASIG_APROBADAS
 - CATEGORIA_BECA_SOCIOECONOMICA
 - TIPO_REGIMEN
 - NUM_MATERIAS_SEGUNDA_MATRICULA
 - NUM_MATERIAS_TERCERA_MATRICULA

Para lograr el acometido de ajustar los tipos de las variables con el fin de que cada una de ellas corresponda a los tipos de datos correctos, se realizó un código dedicado al ajuste de las variables inconsistentes. El código consiste en tres funciones, cada función es específica para tratar las variables inconsistentes de cada reporte, el código será visible a continuación:

Ilustración 57

Función de ajuste de las variables del reporte de Asignaturas Perdidas

```
def ajuste_perdidas(perdidas):
    """
    Ajusta los tipos de variables del reporte de asignaturas perdidas
    """
    for perdida in perdidas:
        for column in perdida.columns:
            if column == "NIVEL_ESTUDIANTE":
                # Se ajusta la columna NIVEL_ESTUDIANTE a object (str)
                perdida[column] = perdida[column].apply(lambda x: "0" + str(x).split(".")[0]
                                                         if not str(x).startswith("0") and
                                                         not str(x).startswith("E") and
                                                         not str(x).startswith("e") and
                                                         not str(x).startswith("n") and
                                                         not str(x).startswith("10")
                                                         else str(x).split(".")[0]
                                                         )
                # Se estandariza Los registros "ELECT", "ELEC" y "EL"
                cambio = {"ELECT" : "Electiva", "ELEC" : "Electiva", "EL" : "Electiva"}
                # Se agregan Los cambios
                perdida[column] = perdida[column].replace(cambio)

            elif column == "NOTA_FINAL" or column == "CRED":
                # Se cambian las variables "NOTA_FINAL" y "CRED" a números de punto flotante (float)
                perdida[column] = perdida[column].replace("[,]", ".", regex=True).replace("[*]", "", regex=True).astype(float)

            elif column == "VECES":
                # Se cambia la variable "VECES" a entero
                perdida[column] = perdida[column].astype(int)

            else:
                # Las demás variables se mantienen o se convierten a object
                perdida[column] = perdida[column].astype(object)
```

Ilustración 58

Función de ajuste de las variables del reporte de Calificaciones por Carrera

```
def ajuste_calificaciones(calificaciones):
    """
    Ajusta los tipos de variables del reporte de asignaturas perdidas
    """
    for calificacion in calificaciones:
        for column in calificacion.columns:
            if column == "nota_parcial" or column == "nota_final" or column == "porcentaje_nota_parcial":
                # Se ajustan las variables a números de punto flotante (float)
                calificacion[column] = calificacion[column].replace("[,]", ".", regex=True).replace("[*]", "", regex=True).astype(float)

            elif column == "VECES":
                # Se ajusta la variable "VECES" a entero
                calificacion[column] = calificacion[column].astype(int)

            else:
                # Las demás variables se mantienen o se convierten a object
                calificacion[column] = calificacion[column].astype(object)
```

Ilustración 59

Función de ajuste de las variables del reporte de Becas

```
def ajuste_becas(becas):  
    '''  
    Ajusta los tipos de variables del reporte de becas  
    '''  
    for beca in becas:  
        for column in beca.columns:  
            if column == "NIVEL_POR_ASIG_APROBADAS":  
                # Se ajusta la variable para ser object  
                beca[column] = beca[column].apply(lambda x: str(x).split(".")[0]  
                                                    if not str(x).startswith("n")  
                                                    else np.nan  
                                                    )  
                beca[column] = beca[column].astype(object)  
  
            elif column == "NUM_MATERIAS_SEGUNDA_MATRICULA" or column == "NUM_MATERIAS_TERCERA_MATRICULA":  
                # Se ajustan las variables para ser números enteros.  
                beca[column] = beca[column].apply(lambda x: str(x).split(".")[0]  
                                                    if not str(x).startswith("n")  
                                                    else np.nan  
                                                    )  
                beca[column] = beca[column].replace([np.nan, np.inf], [-1, 99999]).astype(int)  
  
            else:  
                # Las demás variables se mantienen o se convierten a object  
                beca[column] = beca[column].astype(object)
```

Ilustración 60

Aplicación del ajuste de las variables de los reportes

```
ajuste_perdidas(perdidas)  
ajuste_calificaciones(calificaciones)  
ajuste_becas(becas)
```

A partir las ilustraciones visibles, se puede observar que se realizaron los ajustes de las diferentes variables y sus diferentes tipos de los 3 reportes, en específico, en la última ilustración se visualiza la ejecución de las funciones ajustadoras; a continuación, dado la aplicación del ajuste de las variables de los reportes, es necesario realizar una nueva comparativa, la misma podrá ser contemplada inmediatamente:

Ilustración 61

Resultados de la comparativa tras el ajuste del primer periodo del reporte de Asignaturas Perdidas

```
*****

Analizando el dataframe: perdidas01

IDD                True
PERIODO            True
PLAN_E            True
CAMP              True
DESC_FACULTAD     True
PROGRAMA          True
DESC_PROGRAMA     True
NIVEL_ESTUDIANTE  True
NF               True
MATE              True
CRSE              True
SHRTCKN_CRSE_TITLE True
NOTA_FINAL        True
APROBADO          True
VECES            True
MODO              True
CRED              True
dtype: bool

Características que no poseen los tipos de datos requeridos: []
```

Ilustración 62

Resultados de la comparativa tras el ajuste del segundo periodo del reporte de Asignaturas Perdidas

```
*****

Analizando el dataframe: perdidas02

IDD                True
PERIODO            True
PLAN_E            True
CAMP              True
DESC_FACULTAD     True
PROGRAMA          True
DESC_PROGRAMA     True
NIVEL_ESTUDIANTE  True
NF               True
MATE              True
CRSE              True
SHRTCKN_CRSE_TITLE True
NOTA_FINAL        True
APROBADO          True
VECES            True
MODO              True
CRED              True
dtype: bool

Características que no poseen los tipos de datos requeridos: []
```

Ilustración 63

Resultados de la comparativa tras el ajuste del tercer periodo del reporte de Asignaturas Perdidas

```
*****
Analizando el dataframe: perdidas03

IDD                True
PERIODO            True
PLAN_E            True
CAMP              True
DESC_FACULTAD     True
PROGRAMA          True
DESC_PROGRAMA     True
NIVEL_ESTUDIANTE  True
NF               True
MATE              True
CRSE              True
SHRTCKN_CRSE_TITLE True
NOTA_FINAL        True
APROBADO          True
VECES             True
MODO              True
CRED              True
dtype: bool

Características que no poseen los tipos de datos requeridos: []
```

Ilustración 64

Resultados de la comparativa tras el ajuste del cuarto periodo del reporte de Asignaturas Perdidas

```
*****
Analizando el dataframe: perdidas04

IDD                True
PERIODO            True
PLAN_E            True
CAMP              True
DESC_FACULTAD     True
PROGRAMA          True
DESC_PROGRAMA     True
NIVEL_ESTUDIANTE  True
NF               True
MATE              True
CRSE              True
SHRTCKN_CRSE_TITLE True
NOTA_FINAL        True
APROBADO          True
VECES             True
MODO              True
CRED              True
dtype: bool

Características que no poseen los tipos de datos requeridos: []
```


Ilustración 65

Resultados de la comparativa tras el ajuste del primer periodo del reporte de Calificaciones por Carrera

```
Analizando el dataframe: calificaciones01
```

id_banner_estudiante	True
codigo_periodo	True
desc_periodo	True
codigo_facultad	True
facultad	True
codigo_carrera	True
carrera	True
numero_curso	True
titulo_curso	True
numero_paralelo	True
nombre_parcial	True
descripcion_parcial	True
peso	True
nota_maxima	True
nota_parcial	True
porcentaje_nota_parcial	True
nota_final	True
id_banner_docente	True
codigo_materia	True
descripcion_materia	True
secuencial_parcial	True
fecha_nota_parcial	True
fecha_nota_final	True
codigo_programa	True
descripcion_programa	True
nota_final_SHATCKG	True
cod_razon_de_cambio	True
desc_razon_de_cambio	True
calificacion_del_parcial_score	True
dtype: bool	

```
Características que no poseen los tipos de datos requeridos: []
```

Ilustración 66

Resultados de la comparativa tras el ajuste del segundo periodo del reporte de Calificaciones por Carrera

```
Analizando el dataframe: calificaciones02
```

id_banner_estudiante	True
codigo_perodo	True
desc_perodo	True
codigo_facultad	True
facultad	True
codigo_carrera	True
carrera	True
numero_curso	True
titulo_curso	True
numero_paralelo	True
nombre_parcial	True
descripcion_parcial	True
peso	True
nota_maxima	True
nota_parcial	True
porcentaje_nota_parcial	True
nota_final	True
id_banner_docente	True
codigo_materia	True
descripcion_materia	True
secuencial_parcial	True
fecha_nota_parcial	True
fecha_nota_final	True
codigo_programa	True
descripcion_programa	True
nota_final_SHATCKG	True
cod_razon_de_cambio	True
desc_razon_de_cambio	True
calificacion_del_parcial_score	True
dtype: bool	

```
Características que no poseen los tipos de datos requeridos: []
```

Ilustración 67

Resultados de la comparativa tras el ajuste del tercer periodo del reporte de Calificaciones por Carrera

```
Analizando el dataframe: calificaciones03
```

id_banner_estudiante	True
codigo_perodo	True
desc_perodo	True
codigo_facultad	True
facultad	True
codigo_carrera	True
carrera	True
numero_curso	True
titulo_curso	True
numero_paralelo	True
nombre_parcial	True
descripcion_parcial	True
peso	True
nota_maxima	True
nota_parcial	True
porcentaje_nota_parcial	True
nota_final	True
id_banner_docente	True
codigo_materia	True
descripcion_materia	True
secuencial_parcial	True
fecha_nota_parcial	True
fecha_nota_final	True
codigo_programa	True
descripcion_programa	True
nota_final_SHATCKG	True
cod_razon_de_cambio	True
desc_razon_de_cambio	True
calificacion_del_parcial_score	True
dtype: bool	

```
Características que no poseen los tipos de datos requeridos: []
```

Ilustración 68

Resultados de la comparativa tras el ajuste del cuarto periodo del reporte de Calificaciones por Carrera

Analizando el dataframe: calificaciones04

id_banner_estudiante	True
codigo_periodo	True
desc_periodo	True
codigo_facultad	True
facultad	True
codigo_carrera	True
carrera	True
numero_curso	True
titulo_curso	True
numero_paralelo	True
nombre_parcial	True
descripcion_parcial	True
peso	True
nota_maxima	True
nota_parcial	True
porcentaje_nota_parcial	True
nota_final	True
id_banner_docente	True
codigo_materia	True
descripcion_materia	True
secuencial_parcial	True
fecha_nota_parcial	True
fecha_nota_final	True
codigo_programa	True
descripcion_programa	True
nota_final_SHATCKG	True
cod_razon_de_cambio	True
desc_razon_de_cambio	True
calificacion_del_parcial_score	True
dtype: bool	

Características que no poseen los tipos de datos requeridos: []

Ilustración 69

Resultados de la comparativa tras el ajuste del primer periodo del reporte de Becas

```
*****>
```

```
Analizando el dataframe: becas01
```

ID_BANNER	True
NACIONALIDAD	True
GENERO	True
ETNIA	True
TIPO_COLEGIO	True
PROGRAMA	True
PROGAMA_DESC	True
CAMPUS_PROGRAMA	True
AREA_MATERIAS	True
NIVEL_POR_MATERIAS	True
NIVEL_POR_ASIG_APROBADAS	True
CATEGORIA_BECA_SOCIOECONOMICA	True
TIPO_REGIMEN	True
NUM_MATERIAS_SEGUNDA_MATRICULA	True
NUM_MATERIAS_TERCERA_MATRICULA	True
TIENE_PAGO	True
TIENE_SALDO	True

```
dtype: bool
```

```
Características que no poseen los tipos de datos requeridos: []
```

Ilustración 70

Resultados de la comparativa tras el ajuste del segundo periodo del reporte de Becas

```
*****
```

```
Analizando el dataframe: becas02
```

ID_BANNER	True
NACIONALIDAD	True
GENERO	True
ETNIA	True
TIPO_COLEGIO	True
PROGRAMA	True
PROGAMA_DESC	True
CAMPUS_PROGRAMA	True
AREA_MATERIAS	True
NIVEL_POR_MATERIAS	True
NIVEL_POR_ASIG_APROBADAS	True
CATEGORIA_BECA_SOCIOECONOMICA	True
TIPO_REGIMEN	True
NUM_MATERIAS_SEGUNDA_MATRICULA	True
NUM_MATERIAS_TERCERA_MATRICULA	True
TIENE_PAGO	True
TIENE_SALDO	True

```
dtype: bool
```

```
Características que no poseen los tipos de datos requeridos: []
```

Ilustración 71

Resultados de la comparativa tras el ajuste del tercer periodo del reporte de Becas

```
*****
```

```
Analizando el dataframe: becas03
```

ID_BANNER	True
NACIONALIDAD	True
GENERO	True
ETNIA	True
TIPO_COLEGIO	True
PROGRAMA	True
PROGAMA_DESC	True
CAMPUS_PROGRAMA	True
AREA_MATERIAS	True
NIVEL_POR_MATERIAS	True
NIVEL_POR_ASIG_APROBADAS	True
CATEGORIA_BECA_SOCIOECONOMICA	True
TIPO_REGIMEN	True
NUM_MATERIAS_SEGUNDA_MATRICULA	True
NUM_MATERIAS_TERCERA_MATRICULA	True
TIENE_PAGO	True
TIENE_SALDO	True
dtype:	bool

```
Características que no poseen los tipos de datos requeridos: []
```

Ilustración 72

Resultados de la comparativa tras el ajuste del cuarto periodo del reporte de Becas

```
*****  
  
Analizando el dataframe: becas04  
  
ID_BANNER                True  
NACIONALIDAD              True  
GENERO                    True  
ETNIA                     True  
TIPO_COLEGIO              True  
PROGRAMA                  True  
PROGAMA_DESC              True  
CAMPUS_PROGRAMA           True  
AREA_MATERIAS              True  
NIVEL_POR_MATERIAS         True  
NIVEL_POR_ASIG_APROBADAS   True  
CATEGORIA_BECA_SOCIOECONOMICA True  
TIPO_REGIMEN               True  
NUM_MATERIAS_SEGUNDA_MATRICULA True  
NUM_MATERIAS_TERCERA_MATRICULA True  
TIENE_PAGO                 True  
TIENE_SALDO                True  
dtype: bool  
  
Características que no poseen los tipos de datos requeridos: []
```

En conclusión, se ha logrado exitosamente el ajuste de las variables de los diferentes reportes, lo que significa que los datos se encuentran listos para la identificación de variables en el código, es decir, en consecuencia de haber logrado la distinción en código de los diferentes tipos de datos por cada variable a su correcto estado dado el análisis realizado al comienzo de esta sección, acto seguido, se pueden dividir los marcos de datos correspondientes a cada periodo de cada reporte en marcos de datos de sus variables categóricas y numéricas. Este paso es necesario para llevar a cabo una organizada y correcta exploración de distribuciones y relaciones, para llevar a cabo este procedimiento se realizaron los siguientes códigos con sus respectivos resultados.

Para empezar, se visualizarán los códigos y resultados del reporte de asignaturas perdidas:

Ilustración 73

Código para la identificación de variables numéricas y categóricas del reporte de Asignaturas Perdidas

```
'''
Identificamos los tipos de variables del reporte en código
'''

# Listas que almacenarán los respectivos tipos
numeric_perdidas = []
categoric_perdidas = []

for perdida in perdidas:
    # Identificación de tipos de variables
    numeric = perdida.select_dtypes(include=["float64", "int32"])
    categoric = perdida.select_dtypes(include=["object"])

    # Despliegue de la información
    print("\n" + "*" * 100)
    print(f"\nAnalizando el reporte {perdida.name}...\n")
    print(f"Existen {len(numeric.columns)} variables numericas del reporte, y son: ")
    for variable in numeric.columns:
        print(f"\t- {variable}")
    print(f"\nExisten {len(categoric.columns)} variables categoricas del reporte, y son: ")
    for variable in categoric.columns:
        print(f"\t- {variable}")

    # Se agregan las variables a las listas respectivas
    numeric_perdidas.append(numeric)
    categoric_perdidas.append(categoric)
```

Ilustración 74

Tipos de variables identificadas del primer periodo del reporte de Asignaturas Perdidas

Analizando el reporte perdidas01...

Existen 3 variables numericas del reporte, y son:

- NOTA_FINAL
- VECES
- CRED

Existen 14 variables categoricas del reporte, y son:

- IDD
- PERIODO
- PLAN_E
- CAMP
- DESC_FACULTAD
- PROGRAMA
- DESC_PROGRAMA
- NIVEL_ESTUDIANTE
- NF
- MATE
- CRSE
- SHRTCKN_CRSE_TITLE
- APROBADO
- MODO

Ilustración 75

Tipos de variables identificadas del segundo periodo del reporte de Asignaturas Perdidas

```
*****
```

```
Analizando el reporte perdidas02...
```

```
Existen 3 variables numericas del reporte, y son:
```

- NOTA_FINAL
- VECES
- CRED

```
Existen 14 variables categoricas del reporte, y son:
```

- IDD
- PERIODO
- PLAN_E
- CAMP
- DESC_FACULTAD
- PROGRAMA
- DESC_PROGRAMA
- NIVEL_ESTUDIANTE
- NF
- MATE
- CRSE
- SHRTCKN_CRSE_TITLE
- APROBADO
- MODO

Ilustración 76

Tipos de variables identificadas del tercer periodo del reporte de Asignaturas Perdidas

```
*****
```

```
Analizando el reporte perdidas03...
```

```
Existen 3 variables numericas del reporte, y son:
```

- NOTA_FINAL
- VECES
- CRED

```
Existen 14 variables categoricas del reporte, y son:
```

- IDD
- PERIODO
- PLAN_E
- CAMP
- DESC_FACULTAD
- PROGRAMA
- DESC_PROGRAMA
- NIVEL_ESTUDIANTE
- NF
- MATE
- CRSE
- SHRTCKN_CRSE_TITLE
- APROBADO
- MODO

Ilustración 77

Tipos de variables identificadas del cuarto periodo del reporte de Asignaturas Perdidas

```
*****:
```

```
Analizando el reporte perdidas04...
```

```
Existen 3 variables numericas del reporte, y son:
```

- NOTA_FINAL
- VECES
- CRED

```
Existen 14 variables categoricas del reporte, y son:
```

- IDD
- PERIODO
- PLAN_E
- CAMP
- DESC_FACULTAD
- PROGRAMA
- DESC_PROGRAMA
- NIVEL_ESTUDIANTE
- NF
- MATE
- CRSE
- SHRTCKN_CRSE_TITLE
- APROBADO
- MODO

En consecuencia, de haber finalizado el proceso en el reporte de asignaturas perdidas, a continuación, será visible el procedimiento llevado a cabo sobre el reporte de calificaciones por carrera, tanto su código como resultados:

Ilustración 78

Código para la identificación de variables numéricas y categóricas del reporte de Calificaciones por Carrera

```
'''
Identificamos los tipos de variables del reporte en código
'''

# Listas que almacenarán los respectivos tipos
numeric_calificaciones = []
categoric_calificaciones = []

for calificacion in calificaciones:
    # Identificacion de tipos de variables
    numeric = calificacion.select_dtypes(include=["float64", "int32"])
    categoric = calificacion.select_dtypes(include=["object"])

    # Despliegue de la informacion
    print("\n" + "*" * 100)
    print(f"\nAnalizando el reporte {calificacion.name}...\n")
    print(f"Existen {len(numeric.columns)} variables numericas del reporte, y son: ")
    for variable in numeric.columns:
        print(f"\t- {variable}")
    print(f"\nExisten {len(categoric.columns)} variables categoricas del reporte, y son: ")
    for variable in categoric.columns: print(f"\t- {variable}")

    # Se agregan las variables a las listas respectivas
    numeric_calificaciones.append(numeric)
    categoric_calificaciones.append(categoric)
```

Ilustración 79

Tipos de variables identificadas del primer periodo del reporte de Calificaciones por Carrera

```
*****
```

```
Analizando el reporte calificaciones01...
```

```
Existen 3 variables numericas del reporte, y son:
```

- nota_parcial
- porcentaje_nota_parcial
- nota_final

```
Existen 26 variables categoricas del reporte, y son:
```

- id_banner_estudiante
- codigo_periodo
- desc_periodo
- codigo_facultad
- facultad
- codigo_carrera
- carrera
- numero_curso
- titulo_curso
- numero_paralelo
- nombre_parcial
- descripcion_parcial
- peso
- nota_maxima
- id_banner_docente
- codigo_materia
- descripcion_materia
- secuencial_parcial
- fecha_nota_parcial
- fecha_nota_final
- codigo_programa
- descripcion_programa
- nota_final_SHATCKG
- cod_razon_de_cambio
- desc_razon_de_cambio
- calificacion_del_parcial_score

Ilustración 80

Tipos de variables identificadas del segundo periodo del reporte de Calificaciones por Carrera

```
*****
```

Analizando el reporte calificaciones02...

Existen 3 variables numericas del reporte, y son:

- nota_parcial
- porcentaje_nota_parcial
- nota_final

Existen 26 variables categoricas del reporte, y son:

- id_banner_estudiante
- codigo_periodo
- desc_periodo
- codigo_facultad
- facultad
- codigo_carrera
- carrera
- numero_curso
- titulo_curso
- numero_paralelo
- nombre_parcial
- descripcion_parcial
- peso
- nota_maxima
- id_banner_docente
- codigo_materia
- descripcion_materia
- secuencial_parcial
- fecha_nota_parcial
- fecha_nota_final
- codigo_programa
- descripcion_programa
- nota_final_SHATCKG
- cod_razon_de_cambio
- desc_razon_de_cambio
- calificacion_del_parcial_score

Ilustración 81

Tipos de variables identificadas del tercer periodo del reporte de Calificaciones por Carrera

```
*****:
```

```
Analizando el reporte calificaciones03...
```

```
Existen 3 variables numericas del reporte, y son:
```

- nota_parcial
- porcentaje_nota_parcial
- nota_final

```
Existen 26 variables categoricas del reporte, y son:
```

- id_banner_estudiante
- codigo_periodo
- desc_periodo
- codigo_facultad
- facultad
- codigo_carrera
- carrera
- numero_curso
- titulo_curso
- numero_paralelo
- nombre_parcial
- descripcion_parcial
- peso
- nota_maxima
- id_banner_docente
- codigo_materia
- descripcion_materia
- secuencial_parcial
- fecha_nota_parcial
- fecha_nota_final
- codigo_programa
- descripcion_programa
- nota_final_SHATCKG
- cod_razon_de_cambio
- desc_razon_de_cambio
- calificacion_del_parcial_score

Ilustración 82

Tipos de variables identificadas del cuarto periodo del reporte de Calificaciones por Carrera

```
*****:
```

```
Analizando el reporte calificaciones04...
```

```
Existen 3 variables numericas del reporte, y son:
```

- nota_parcial
- porcentaje_nota_parcial
- nota_final

```
Existen 26 variables categoricas del reporte, y son:
```

- id_banner_estudiante
- codigo_periodo
- desc_periodo
- codigo_facultad
- facultad
- codigo_carrera
- carrera
- numero_curso
- titulo_curso
- numero_paralelo
- nombre_parcial
- descripcion_parcial
- peso
- nota_maxima
- id_banner_docente
- codigo_materia
- descripcion_materia
- secuencial_parcial
- fecha_nota_parcial
- fecha_nota_final
- codigo_programa
- descripcion_programa
- nota_final_SHATCKG
- cod_razon_de_cambio
- desc_razon_de_cambio
- calificacion_del_parcial_score

Finalmente, tras haber realizado el proceso sobre los reportes de asignaturas perdidas y calificaciones por carrera, ha llegado el momento de identificar las variables del reporte de becas, por lo que, su código como resultados podrán ser observados a continuación:

Ilustración 83

Código para la identificación de variables numéricas y categóricas del reporte de Becas

```
...
Identificamos los tipos de variables del reporte en código
...

# Listas que almacenarán los respectivos tipos
numeric_becas = []
categoric_becas = []

for beca in becas:
    # Identificación de tipos de variables
    numeric = beca.select_dtypes(include=["float64", "int32"])
    categoric = beca.select_dtypes(include=["object"])

    # Despliegue de la información
    print("\n" + "*" * 100)
    print(f"\nAnalizando el reporte {beca.name}...\n")
    print(f"Existen {len(numeric.columns)} variables numericas del reporte, y son: ")
    for variable in numeric.columns:
        print(f"\t- {variable}")
    print(f"\nExisten {len(categoric.columns)} variables categoricas del reporte, y son: ")
    for variable in categoric.columns: print(f"\t- {variable}")

    # Se agregan las variables a las listas respectivas
    numeric_becas.append(numeric)
    categoric_becas.append(categoric)
```

Ilustración 84

Tipos de variables identificadas del primero periodo del reporte de Becas

```
*****;
```

```
Analizando el reporte becas01...
```

```
Existen 2 variables numericas del reporte, y son:
```

- NUM_MATERIAS_SEGUNDA_MATRICULA
- NUM_MATERIAS_TERCERA_MATRICULA

```
Existen 15 variables categoricas del reporte, y son:
```

- ID_BANNER
- NACIONALIDAD
- GENERO
- ETNIA
- TIPO_COLEGIO
- PROGRAMA
- PROGRAM_DESC
- CAMPUS_PROGRAMA
- AREA_MATERIAS
- NIVEL_POR_MATERIAS
- NIVEL_POR_ASIG_APROBADAS
- CATEGORIA_BECA_SOCIOECONOMICA
- TIPO_REGIMEN
- TIENE_PAGO
- TIENE_SALDO

Ilustración 85

Tipos de variables identificadas del segundo periodo del reporte de Becas

Analizando el reporte becas02...

Existen 2 variables numericas del reporte, y son:

- NUM_MATERIAS_SEGUNDA_MATRICULA
- NUM_MATERIAS_TERCERA_MATRICULA

Existen 15 variables categoricas del reporte, y son:

- ID_BANNER
- NACIONALIDAD
- GENERO
- ETNIA
- TIPO_COLEGIO
- PROGRAMA
- PROGAMA_DESC
- CAMPUS_PROGRAMA
- AREA_MATERIAS
- NIVEL_POR_MATERIAS
- NIVEL_POR_ASIG_APROBADAS
- CATEGORIA_BECA_SOCIOECONOMICA
- TIPO_REGIMEN
- TIENE_PAGO
- TIENE_SALDO

Ilustración 86

Tipos de variables identificadas del tercer periodo del reporte de Becas

Analizando el reporte becas03...

Existen 2 variables numericas del reporte, y son:

- NUM_MATERIAS_SEGUNDA_MATRICULA
- NUM_MATERIAS_TERCERA_MATRICULA

Existen 15 variables categoricas del reporte, y son:

- ID_BANNER
- NACIONALIDAD
- GENERO
- ETNIA
- TIPO_COLEGIO
- PROGRAMA
- PROGAMA_DESC
- CAMPUS_PROGRAMA
- AREA_MATERIAS
- NIVEL_POR_MATERIAS
- NIVEL_POR_ASIG_APROBADAS
- CATEGORIA_BECA_SOCIOECONOMICA
- TIPO_REGIMEN
- TIENE_PAGO
- TIENE_SALDO

Ilustración 87

Tipos de variables identificadas del cuarto periodo del reporte de Becas

```
*****  
  
Analizando el reporte becas04...  
  
Existen 2 variables numericas del reporte, y son:  
- NUM_MATERIAS_SEGUNDA_MATRICULA  
- NUM_MATERIAS_TERCERA_MATRICULA  
  
Existen 15 variables categoricas del reporte, y son:  
- ID_BANNER  
- NACIONALIDAD  
- GENERO  
- ETNIA  
- TIPO_COLEGIO  
- PROGRAMA  
- PROGAMA_DESC  
- CAMPUS_PROGRAMA  
- AREA_MATERIAS  
- NIVEL_POR_MATERIAS  
- NIVEL_POR_ASIG_APROBADAS  
- CATEGORIA_BECA_SOCIOECONOMICA  
- TIPO_REGIMEN  
- TIENE_PAGO  
- TIENE_SALDO
```

En síntesis, se puede observar que se han identificado correctamente los tipos de variables de los distintos periodos de sus respectivos reportes, esto se puede visualizar en las ilustraciones correspondientes a los tipos de variables identificados dado el periodo y el reporte; donde para todos los casos de cada reporte concuerda con los resultados del análisis realizado al comienzo de esta sección.

4.1.9. Selección preliminar de variables relevantes

Hasta este momento se ha realizado una correcta anonimización de datos y una apropiada identificación de los tipos de variables disponibles de cada reporte, en esta ocasión, se llevará a cabo una selección preliminar de variables relevantes. No obstante, esta acción muy probablemente pueda generar controversia en el ilustre lector, debido a que la selección de

variables relevantes usualmente se la realiza después de un debido proceso de exploración de los datos, sin embargo, dada la inmediata siguiente explicación, esta acción poseerá toda la coherencia imaginable.

Para comenzar la defensa y argumentación del porqué de esta sección, se tiene que recordar que hasta este momento se han analizado varias características de las variables que han sido utilizadas en la presente investigación, esto, ha permitido determinar cierto grado de relevancia para cada una de las variables de cada reporte; sin embargo, aunque es necesario realizar una exploración más detallada para determinar la utilidad real de las variables más allá de meras conjeturas, también es cierto que algunas variables no tienen mayor relevancia a simple vista, por ejemplo, aquellas variables que poseen un único valor o aquellas con miles de valores pero que no son útiles en tal magnitud y podrían ser reestructuradas de mejor manera, etc.

A partir de lo mencionado, a continuación, se presentará una lista de los motivos por los cuales una variable será considerada como irrelevante:

- 1) **Variables con un único valor y sin relevancia:** Estas variables no aportan información significativa para la investigación, debido a que todos sus registros contienen el mismo valor, y, además de ello, no presentan utilidad con respecto a su existencia en los demás reportes. Las variables que pertenecen a este contexto se detallan a continuación.

Tabla 12
Variables con un único valor y sin relevancia

Reporte	Variable	Razón
Asignaturas Per- didas	NF	Debido a que todos los datos de todos los reportes pertenecen a pregrado.
Asignaturas Per- didas	MODO	Todas las calificaciones pertenecen a la misma escala.
Asignaturas Per- didas	CAMP	Todos los registros de todos los reportes son del campus en Quito.

Calificaciones por Carrera	nota_maxima	Representa la nota máxima posible, dado que la escala de calificación es la misma en todos los periodos y reportes, no representa utilidad.
Calificaciones por Carrera	calificacion_del_parcial_score	Representa la nota máxima posible, dado que la escala de calificación es la misma en todos los periodos y reportes, no representa utilidad.
Becas	TIPO_REGIMEN	Todos los registros corresponden al mismo tipo política con el que se fue asignada la beca socioeconómica.
Becas	CAMPUS_PROGRAMA	Todos los registros de todos los reportes son del campus en Quito.

2) Variables con miles de valores y sin utilidad para el análisis: Estas variables contienen una gran cantidad de valores únicos que dificultan el análisis y no aportan información valiosa para el objetivo de la investigación. El planteamiento sugiere que, en lugar de enfocarse en cada valor individual, es más práctico considerar una representación más generalizada de estas variables. Las variables que se aplican a este contexto son aquellas que representan a las asignaturas y se debe a las siguientes razones:

- a. Gran cantidad de valores:** Las variables correspondientes a las asignaturas contienen miles de valores diferentes debido a la diversidad de asignaturas de la Pontificia Universidad Católica del Ecuador, esta gran cantidad de valores dificulta el análisis y puede generar complejidad adicional al modelo, por lo que, en lugar de enfocarse en cada asignatura individualmente, es más práctico y eficiente considerar una variable que represente el área de asignatura de manera más generalizada.

- b. Simplificación y generalización:** Al agrupar las asignaturas en categorías más amplias, como áreas de asignatura, se logra una simplificación y generalización de los datos, lo que implica la reducción de la complejidad y el ruido en el modelo, permitiendo un enfoque más claro en las características relevantes para la predicción del rendimiento académico.

Tabla 13

Variables con miles de valores y sin utilidad para el análisis

Reporte	Variable
Asignaturas Perdidas	CRSE
Asignaturas Perdidas	SHRTCKN_CRSE_TITLE
Calificaciones por Carrera	numero_curso
Calificaciones por Carrera	titulo_curso

- 3) Variables redundantes:** Estas variables se refieren a los mismos datos, es decir, aquellas variables que son otras variables, pero en código o viceversa, por ejemplo, una variable que especifica una carrera y otra que especifica el código de la misma carrera. Las variables que pertenecen a este contexto se mostrarán en la tabla siguiente.

Tabla 14

Variables redundantes

Reporte	Variable	Razón
Asignaturas Perdidas	PROGRAMA	Ya se conoce el nombre de la carrera.
Calificaciones por Carrera	desc_periodo	Ya se conoce el identificador del periodo.

Calificaciones por Carrera	codigo_facultad	Ya se conoce el nombre de las facultades.
Calificaciones por Carrera	codigo_carrera	Ya se conoce el nombre de las facultades.
Calificaciones por Carrera	descripcion_parcial	Ya se dispone de un identificador de parciales.
Calificaciones por Carrera	secuencial_parcial	Debido a que menciona la información del identificador de los parciales.
Calificaciones por Carrera	porcentaje_nota_parcial	Representa el mismo valor que el peso.
Calificaciones por Carrera	codigo_materia	Ya se dispone de los nombres de las asignaturas.
Calificaciones por Carrera	codigo_programa	Ya se dispone de los nombres de las carreras.
Calificaciones por Carrera	cod_razon_de_cambio	Ya se dispone del valor real de la razón del cambio de la nota final.
Calificaciones por Carrera	numero_paralelo	El registro realmente resulta redundante debido a que la distinción de un paralelo y otro es el docente que imparte la asignatura respectiva por lo que resulta innecesaria ya que ya se dispone de esta información.

- 4) Variables que a pesar de no tener un análisis profundo no son relevantes:** Estas variables no presentan ningún tipo de valor agregado a la investigación sin necesidad de realizar un análisis profundo, a continuación, se podrán visualizar las variables que pertenecen a este contexto.

Tabla 15

Variables que a pesar de no tener un análisis profundo no son relevantes

Reporte	Variable	Razón
Calificaciones por Carrera	carrera	Debido a que es la unión entre la variable "facultad" y "programa", podría ser incluso considerada redundante.
Calificaciones por Carrera	fecha_nota_parcial	Debido a que únicamente representa cuando se subió la nota parcial al sistema Banner.
Calificaciones por Carrera	fecha_nota_final	Debido a que únicamente representa cuando se subió la nota final al sistema Banner.
Calificaciones por Carrera	desc_razon_de_cambio	Debido a que se refiere únicamente a la razón de ajuste de nota final

- 5) **Variables que, dado el enfoque de la investigación, son irrelevantes:** Estas variables podrían ser mayoritariamente útiles en multitud de contextos referidas a la investigación del rendimiento académico, sin embargo, no resultan relevantes dada la investigación actual. Estas serán detalladas en la siguiente tabla.

Tabla 16

Variables que, dado el enfoque de la investigación, son irrelevantes

Reporte	Variables
Calificaciones por Carrera	nombre_parcial
Calificaciones por Carrera	peso
Calificaciones por Carrera	nota_parcial

En efecto, usted estimado lector, ha leído bien, se han considerado las variables relacionadas a todos los parciales de un periodo académico dado y sus calificaciones como irrelevantes, y, pues sí, nuevamente esta sección genera controversia pues aparentemente estas variables deberían tener un impacto significativo en el rendimiento de un estudiante determinado.

Sin embargo, es importante entender que el enfoque de la investigación es predecir si un estudiante aprobará una asignatura sin haberla cursado o estando por cursarla, en lugar de predecir el rendimiento académico una vez que se está cursando la asignatura; en otras palabras, las variables relacionadas con las calificaciones parciales serían útiles si se estuviera evaluando la predicción del rendimiento académico en cada parcial, es decir, determinar si, dado el resultado de un parcial y otros factores, a un estudiante determinado le irá mejor o peor en el próximo parcial. Sin embargo, la investigación actual se centra en predecir si un estudiante aprobará o reprobará una asignatura sin haberla tomado o estando por tomarla, es por esta razón que se ha tomado la decisión de eliminar estas variables.

No obstante, la variable correspondiente a *nota_parcial* no se eliminará hasta la siguiente etapa de selección de variables relevantes, ya que, podría ser interesante realizar un análisis de su distribución. Por otra parte, siguiendo la misma lógica realmente tampoco sería útil la variable *nota_final*, sin embargo, teniendo en cuenta que en primer lugar se desea evaluar el rendimiento académico, es necesario contar con variables que esclarecen más evidentemente el rendimiento que un estudiante tuvo en una asignatura dada, es decir, en consecuencia de que se utilizará un enfoque para evaluar el rendimiento académico en base al resultado final que determinados estudiantes tuvieron en alguna asignatura en términos concretos del rendimiento final y no progresivo, se utilizará la variable *nota_final* y no los datos de la variable *nota_parcial* para la selección de variables que influyen directamente en el desempeño académico.

Por lo tanto, las variables que han sido identificadas como irrelevantes dado los diferentes enfoques especificados son las visibles en la siguiente tabla:

Tabla 17
Variables seleccionadas como irrelevantes

Reporte	Variable	Reemplazada por	Razón
---------	----------	-----------------	-------

Asignaturas Perdidas	CRSE	MATE	Gran cantidad de valores
Asignaturas Perdidas	SHRTCKN_CRSE_TITLE	MATE	Gran cantidad de valores
Asignaturas Perdidas	NF	Sin reemplazo	Único valor irrelevante
Asignaturas Perdidas	MODO	Sin reemplazo	Único valor irrelevante
Asignaturas Perdidas	CAMP	Sin reemplazo	Único valor irrelevante
Asignaturas Perdidas	PROGRAMA	SHRTCKN_CRSE_TIT- TLE	Variable re- dundante
Calificaciones por Carrera	numero_curso	descripcion_materia	Gran cantidad de valores
Calificaciones por Carrera	titulo_curso	descripcion_materia	Gran cantidad de valores
Calificaciones por Carrera	nota_maxima	Sin reemplazo	Único valor irrelevante
Calificaciones por Carrera	calificacion_del_parcial_score	Sin reemplazo	Único valor irrelevante
Calificaciones por Carrera	desc_periodo	codigo_periodo	Variable re- dundante
Calificaciones por Carrera	codigo_facultad	facultad	Variable re- dundante
Calificaciones por Carrera	codigo_carrera	carrera	Variable re- dundante

Calificaciones por Carrera	descripcion_parcial	nombre_parcial	Variable redundante
Calificaciones por Carrera	codigo_materia	descripcion_materia	Variable redundante
Calificaciones por Carrera	porcentaje_nota_parcial	nota_parcial	Variable redundante
Calificaciones por Carrera	codigo_programa	descripcion_programa	Variable redundante
Calificaciones por Carrera	cod_razon_de_cambio	desc_razon_de_cambio	Variable redundante
Calificaciones por Carrera	numero_paralelo	id_banner_docente	Variable redundante
Calificaciones por Carrera	secuencial_parcial	nombre_parcial	Variable redundante
Calificaciones por Carrera	carrera	Sin reemplazo	No relevante sin necesidad de análisis
Calificaciones por Carrera	fecha_nota_parcial	Sin reemplazo	No relevante sin necesidad de análisis
Calificaciones por Carrera	fecha_nota_final	Sin reemplazo	No relevante sin necesidad de análisis
Calificaciones por Carrera	desc_razon_de_cambio	Sin reemplazo	No relevante sin necesidad de análisis

Calificaciones por Carrera	nombre_parcial	Sin reemplazo	No relevante dado el enfoque de la investigación
Calificaciones por Carrera	peso	Sin reemplazo	No relevante dado el enfoque de la investigación
Becas	TIPO_REGIMEN	Sin reemplazo	Único valor irrelevante
Becas	CAMPUS_PROGRAMA	Sin reemplazo	Único valor irrelevante
Becas	PROGRAMA	PROGAMA_DESC	Variable redundante

Como se puede observar, se han seleccionado 29 de las 31 variables consideradas irrelevantes, esto significa que las variables `nota_parcial` y `nota_final` aun no serán eliminadas; sin embargo, las demás 29 variables irrelevantes serán sujetas a procesos de eliminación a partir de los siguientes códigos:

Ilustración 88

Código de eliminación de variables irrelevantes del reporte de Asignaturas Perdidas

```
columnas_por_eliminar = ["PROGRAMA", "CRSE", "SHRTCKN_CRSE_TITLE", "CAMP", "NF", "MOD0", "CRSE"]
for perdida in perdidas:
    perdida.drop(columnas_por_eliminar, axis=1, inplace=True)
```

```
# Vista previa de la eliminación
perdidas[0].head()
```


Ilustración 89

Código de la eliminación de variables irrelevantes del reporte de Calificaciones por Carrera

```
columnas_por_eliminar = ["numero_curso", "titulo_curso", "nota_maxima",  
                        "porcentaje_nota_parcial", "calificacion_del_parcial_score", "desc_periodo",  
                        "codigo_facultad", "codigo_carrera", "descripcion_parcial",  
                        "codigo_materia", "codigo_programa", "cod_razon_de_cambio",  
                        "carrera", "fecha_nota_parcial", "fecha_nota_final",  
                        "desc_razon_de_cambio", "numero_paralelo", "secuencial_parcial",  
                        "nombre_parcial", "peso"]  
  
for calificacion in calificaciones:  
    calificacion.drop(columnas_por_eliminar, axis=1, inplace=True)
```

```
# Vista previa de la eliminación  
calificaciones[0].head()
```

Ilustración 90

Código de la eliminación y ajuste de variables irrelevantes del reporte de Becas

```
# Eliminación de las variables irrelevantes  
columnas_por_eliminar = ["PROGRAMA", "TIPO_REGIMEN", "CAMPUS_PROGRAMA"]  
for beca in becas:  
    beca.drop(columnas_por_eliminar, axis=1, inplace=True)
```

```
# Ajuste de la variable AREA_MATERIAS  
print("Actual formato del contenido de la variable AREA_MATERIAS: ", becas[0]["AREA_MATERIAS"].unique()[0])  
for beca in becas:  
    beca["AREA_MATERIAS"] = beca["AREA_MATERIAS"].apply(lambda x: str(x)[-2:])  
print("Actual formato del contenido de la variable AREA_MATERIAS: ", becas[0]["AREA_MATERIAS"].unique()[0])
```

```
Actual formato del contenido de la variable AREA_MATERIAS:  Q148-01  
Actual formato del contenido de la variable AREA_MATERIAS:  01
```

Una vez realizada la ejecución de los códigos acabados de presentar, se ha concretado la eliminación de 29 variables de los distintos reportes, esto dio como resultado las ilustraciones de los marcos de datos en código que se podrán visualizar seguido a este párrafo, no obstante, es necesario mencionar que únicamente se evidenciarán ilustraciones de los primeros periodos de cada reporte, esto, debido a que la eliminación se ha llevado a cabo sobre todos los periodos, y, ya que todos los periodos constan de las mismas variables, resulta innecesario desplegar todos los marcos de datos de cada periodo si la finalidad es la demostración de la ejecución de la eliminación de variables irrelevantes.

Ilustración 91

Vista del marco de datos del reporte de Asignaturas Perdidas

```
# Vista previa de la eliminación
perdidas[0].head()
```

	IDD	PERIODO	PLAN_E	DESC_FACULTAD	DESC_PROGRAMA	NIVEL_ESTUDIANTE	MATE	NOTA_FINAL	APROBADO	VECES	CRED
0	E5214	202161	2	Facultad Psicología	Psicología Clínica	09	9	28.0	N	1	6.0
1	E4739	202161	3	Facultad Cs Humanas	Sociología Política	04	3	11.0	N	1	3.0
2	E10373	202161	1	Facultad Cs Adm. y Contables	Negocios Internacionales	04	4	27.0	N	1	3.0
3	E2887	202161	3	Facultad Economía	Economía	04	3	23.0	N	1	4.0
4	E2887	202161	3	Facultad Economía	Economía	04	3	28.0	N	1	4.0

Ilustración 92

Vista del marco de datos del reporte de Calificaciones por Carrera

```
# Vista previa de la eliminación
calificaciones[0].head()
```

	id_banner_estudiante	codigo_periodes	facultad	nota_parcial	nota_final	id_banner_docente	descripcion_materia	descripcion_programa	nota_final_SHATCKG
0	E8695	202161	Facultad Cs Adm. y Contables	35.93	35.68	D936	Cien.Natur.Matemát.y Estadíst.	Administración de Empresas	36
1	E8695	202161	Facultad Cs Adm. y Contables	41.08	35.68	D936	Cien.Natur.Matemát.y Estadíst.	Administración de Empresas	36
2	E8695	202161	Facultad Cs Adm. y Contables	28.38	35.68	D936	Cien.Natur.Matemát.y Estadíst.	Administración de Empresas	36
3	E8695	202161	Facultad Cs Adm. y Contables	37.33	35.68	D936	Cien.Natur.Matemát.y Estadíst.	Administración de Empresas	36
4	E8695	202161	Facultad Cs Adm. y Contables	34.81	34.79	D619	Administración	Administración de Empresas	35

Ilustración 93

Parte 1 de la vista del marco de datos del reporte de Becas

```
# Vista previa de la eliminación
becas[0].head()
```

	ID_BANNER	NACIONALIDAD	GENERO	ETNIA	TIPO_COLEGIO	PROGAMA_DESC	AREA_MATERIAS	NIVEL_POR_MATERIAS
0	E9532	Ecuatoriano	Mujer	NaN	NaN	Enfermería	01	01
1	E2371	Ecuatoriano	Hombre	Mestizo	NaN	Economía	07	07
2	E2838	Ecuatoriano	Mujer	NaN	Part L	Negocios Internacionales	04	04
3	E3781	Ecuatoriano	Mujer	Mestizo	NaN	Arquitectura	10	10
4	E2249	Ecuatoriano	Hombre	NaN	NaN	Teología	02	02

Ilustración 94

Parte 2 de la vista del marco de datos del reporte de Becas

```
# Vista previa de La eliminación
becas[0].head()
```

IAS	NIVEL_POR_ASIG_APROBADAS	CATEGORIA_BECA_SOCIOECONOMICA	NUM_MATERIAS_SEGUNDA_MATRICULA	NUM_MATERIAS_TERCERA_MATRICULA	TIENE_PAGO	TIENE_SALDO
01	2	A	1	0	SI	NO
07	8	C	0	0	SI	NO
04	7	C	1	0	SI	NO
10	10	A	0	0	SI	NO
02	2	B	0	0	SI	NO

Hasta este momento, se ha demostrado una impactante reducción de la cantidad de variables que en un principio se poseía, sin embargo, dado el contexto con el que se tomaron las respectivas decisiones que han permitido alcanzar este punto del trabajo de titulación, esta reducción de variables no resulta sinónimo de una eventualidad negativa, sino, un paso en el proceso de poder seleccionar las mejores variables para el modelo de predicción de rendimiento académico.

En contraste con lo realizado hasta el momento, es de vital importancia mencionar que se volvió a realizar un proceso de identificación y selección de tipos de variables en código, esto se debe a que, tras la eliminación de las variables consideradas irrelevantes, no se han efectuado estos cambios en las listas que contenían las variables numéricas y categóricas de cada periodo de cada reporte respectivamente. Sus resultados podrán ser evaluados en las ilustraciones consiguientes, sin embargo, se debe tener en cuenta que únicamente se manifestarán los resultados de los códigos y no estos, puesto que son los mismos a los realizados en la sección anterior, además, se mostrará únicamente los resultados del primer periodo de cada reporte ya que el mismo concepto aplica para todos demás periodos dado el reporte.

Ilustración 95

Identificación en código de tipos de variables relevantes del reporte de Asignaturas Perdidas

```
*****  
  
Analizando el reporte perdidas01...  
  
Existen 3 variables numericas del reporte, y son:  
- NOTA_FINAL  
- VECES  
- CRED  
  
Existen 8 variables categoricas del reporte, y son:  
- IDD  
- PERIODO  
- PLAN_E  
- DESC_FACULTAD  
- DESC_PROGRAMA  
- NIVEL_ESTUDIANTE  
- MATE  
- APROBADO
```

Ilustración 96

Identificación en código de tipos de variables relevantes del reporte de Calificaciones por Carrera

```
*****  
  
Analizando el reporte calificaciones01...  
  
Existen 2 variables numericas del reporte, y son:  
- nota_parcial  
- nota_final  
  
Existen 7 variables categoricas del reporte, y son:  
- id_banner_estudiante  
- codigo_periodo  
- facultad  
- id_banner_docente  
- descripcion_materia  
- descripcion_programa  
- nota_final_SHATCKG
```

Ilustración 97

Identificación en código de tipos de variables relevantes del reporte de Becas

```
*****:

Analizando el reporte becas01...

Existen 2 variables numericas del reporte, y son:
- NUM_MATERIAS_SEGUNDA_MATRICULA
- NUM_MATERIAS_TERCERA_MATRICULA

Existen 12 variables categoricas del reporte, y son:
- ID_BANNER
- NACIONALIDAD
- GENERO
- ETNIA
- TIPO_COLEGIO
- PROGAMA_DESC
- AREA_MATERIAS
- NIVEL_POR_MATERIAS
- NIVEL_POR_ASIG_APROBADAS
- CATEGORIA_BECA_SOCIOECONOMICA
- TIENE_PAGO
- TIENE_SALDO
```

En conclusión, de la sección, se ha logrado la identificación acertada de variables irrelevantes y aquellas que aun necesitan de un mayor y enfocado análisis para la determinación de su relevancia, es por lo que, en las siguientes secciones se explorará a detalle cada una de las variables tanto numéricas como categóricas de cada periodo de cada reporte de datos, y, después, se explorarán las relaciones que estas variables poseen con respecto al rendimiento académico.

4.1.10. Exploración de variables numéricas

Como se menciona por Francisco Barón López, “Cuando se aborda el estudio de un conjunto de datos, antes de introducimos en cuestiones más detalladas, es necesario hacer una exploración inicial de los mismos. Así se puede tener una idea más clara de las características principales de los datos” (López, 2022). Esto quiere decir, que es necesario realizar una exploración de las variables con las que se dispone por cada reporte con la finalidad de

aproximarse a seleccionar las variables ideales para el modelo de predicción de rendimiento académico.

En primer lugar, entendiendo que se le realizará una exploración de las variables numéricas, “la distinción entre datos discretos y continuos es importante para decidir qué método de análisis estadístico utilizar” (Orellana, 2001, pág. 6). Sin embargo, esta distinción de los tipos de datos ya fue previamente realizada en anteriores secciones en términos de la identificación de variables numéricas de tipo discretos y continuos.

En esta sección se realizará la exploración de las variables numéricas por cada reporte, entre los análisis exploratorios que se realizarán, en primer lugar, se analizarán las estadísticas descriptivas debido a que “Descriptive statistics include those that summarize the central tendency, dispersión and shape of the dataset’s distribution, excluding NaN values.” [Las estadísticas descriptivas incluyen aquellas que resumen la tendencia central, la dispersión y la forma de la distribución del conjunto de datos, excluyendo los valores nulos.] (pandas, s. f.). Lo que implica que servirá para detectar valores atípicos y distribuciones anómalas, adicionalmente, se utilizará histogramas, ya que son gráficos que “ayudan a ver el centro, la extensión y la forma de un conjunto de datos. (...) Se pueden usar para comprobar valores extremos o atípicos y ayudar a comprender la distribución de sus datos.” (jmp, s. f.). Finalmente, se utilizarán diagramas de caja y bigotes debido a que “son una presentación visual que describe varias características importantes, al mismo tiempo, tales como la dispersión y simetría” (Espejo, s. f., pág. 1), estas gráficas son especialmente buenas para detectar valores atípicos.

Como análisis adicional, también se realizarán gráficos de densidad ya que “visualiza la distribución de datos cuantitativos en un intervalo o período de tiempo continuo” (Umbarilla, 2022), lo que resulta ideal considerando que se consta de registros que siguen un periodo de tiempo continuo entre periodos académico.

Para empezar la exploración de variables numéricas, se realizará sobre el reporte de asignaturas perdidas con un análisis de sus estadísticas descriptivas, pues, como ya se ha mencionado, gracias a esta técnica se presentan las medidas de tendencia central, dispersión y forma de cada variable numérica. A continuación, progresivamente que irán presentando las estadísticas descriptivas correspondientes a cada periodo del reporte seguido de sus observaciones:

Se comenzará con la visualización de las estadísticas descriptivas del primer periodo:

Ilustración 98

Estadísticas descriptivas del primer periodo del reporte de Asignaturas Perdidas

```
numeric_perdidas[0].describe()
```

	NOTA_FINAL	VECES	CRED
count	1670.000000	1670.000000	1670.000000
mean	18.171257	1.203593	3.788623
std	10.014221	0.463644	1.826242
min	0.000000	1.000000	0.000000
25%	10.000000	1.000000	3.000000
50%	21.000000	1.000000	3.000000
75%	26.000000	1.000000	4.500000
max	49.000000	4.000000	12.500000

Las observaciones destacables de las estadísticas descriptivas de este periodo son las siguientes:

- Existen 1670 registros totales por cada variable numérica, lo que podría indicar que no existen valores nulos.
- El promedio de notas finales de las asignaturas repetidas es de 18.17, lo que resulta relativamente bajo.
- En promedio, las asignaturas que se perdieron son repetidas 1.2 veces.
- En promedio, las asignaturas perdidas tienen un valor de 3.8 créditos.
- Dada la desviación estándar, es factible inferir que la mayor parte de las notas se encontrarán entre 8.16 y 28.18.

- Dada la desviación estándar, se puede inferir que la mayor parte de registros referentes a la cantidad de veces que se repite una asignatura se da entre 0.74 y 1.67.
- Dada la desviación estándar, se entiende que la mayor parte de registros referentes a los créditos que valen cada asignatura son alrededor de 1.96 y 5.61.
- Se puede observar que en principio parecen no existir valores atípicos en los registros de las notas finales, teniendo como nota mínima 0 y máxima 49.
- Dada la observación de la cantidad mínima de repeticiones de una asignatura, se pueden dar dos suposiciones:
- Se registra como repetición una vez que se pierde la asignatura.
- Todas las asignaturas que se perdieron en este periodo son repetidas.
- Se puede observar que el 75% de las notas se encuentran por debajo o igual a 26, lo que podría significar que la nota máxima encontrada de 49 podría ser un valor atípico.
- Igualmente, se aplica para las repeticiones debido a que el 75% de sus datos son iguales a 1 y después, se tiene un valor máximo de 4 repeticiones, lo que resulta como más resaltante la posibilidad de que este valor sea un valor atípico debido a que, dadas las políticas de la universidad, no se puede repetir una sola asignatura más de 3 veces; sin embargo, dadas las observaciones hechas para la cantidad mínima de repeticiones, podría no considerarse como un valor atípico, sino como un referente de que el estudiante reprobó su tercera matrícula.

- Finalmente, se observa que igualmente es probable que existan valores atípicos en el registro de créditos debido a que el 75% de sus registros poseen valores menores o iguales a 4.5 y se tiene un valor máximo de 12.5.

Se continuará con la visualización de las estadísticas descriptivas del segundo periodo:

Ilustración 99

Estadísticas descriptivas del segundo periodo del reporte de Asignaturas Perdidas

```
numeric_perdidas[1].describe()
```

	NOTA_FINAL	VECES	CRED
count	2104.000000	2104.000000	2104.000000
mean	18.524606	1.216730	3.729800
std	10.536267	0.493982	1.990783
min	0.000000	1.000000	0.000000
25%	10.000000	1.000000	3.000000
50%	22.000000	1.000000	3.000000
75%	27.000000	1.000000	4.500000
max	39.000000	4.000000	20.000000

Las observaciones destacables de las estadísticas descriptivas de este periodo son las siguientes:

- Existen 2104 registros totales por cada variable numérica, lo que podría indicar que no existen valores nulos. Además, se puede observar que muy presumiblemente hubo más estudiantes reprobados en este semestre en comparación con el anterior, específicamente 434 estudiantes más reprobados.

- El promedio de notas finales de las asignaturas repetidas es de 18.52, lo que resulta relativamente bajo pero superior por 0.35 al anterior periodo.
- En promedio, las asignaturas que se perdieron son repetidas 1.2 veces, igual que en el periodo anterior.
- En promedio, las asignaturas perdidas tienen un valor de 3.7 créditos, variando únicamente una décima del periodo anterior.
- Dada la desviación estándar, se está en condiciones de inferir que la mayor parte de las notas se encontrarán entre 8 y 29.06.
- Dada la desviación estándar, es factible inferir que la mayor parte de registros referentes a la cantidad de veces que se repite una asignatura se da entre 0.72 y 1.71.
- Dada la desviación estándar, se posible inferir que la mayor parte de registros referentes a los créditos que valen cada asignatura son alrededor de 1.74 y 5.72.
- Se puede observar que en principio parecen no existir valores atípicos en los registros de las notas finales, teniendo como nota mínima 0 y máxima 39.
- Dada la observación de la cantidad mínima de repeticiones de una asignatura, se pueden dar dos suposiciones:
 - Se registra como repetición una vez que se pierde la asignatura.
 - Todas las asignaturas que se perdieron en este periodo son repetidas.

- Se puede observar que el 75% de las notas se encuentran por debajo o igual a 27, sin embargo, al tener un valor máximo de 39, realmente no necesariamente se refiere a un valor atípico.
- Igualmente, se aplica para las repeticiones debido a que el 75% de sus datos son iguales a 1 y después, se tiene un valor máximo de 4 repeticiones, lo que resulta como más resaltante la posibilidad de que este valor sea un valor atípico debido a que, dadas las políticas de la universidad, no se puede repetir una sola asignatura más de 3 veces; sin embargo, dadas las observaciones hechas para la cantidad mínima de repeticiones, podría no considerarse como un valor atípico, sino como un referente de que el estudiante reprobó su tercera matrícula.
- Finalmente, es observable que igualmente pueden existir valores atípicos en el registro de créditos debido a que el 75% de sus registros poseen valores menores o iguales a 4.5 y se tiene un valor máximo de 20.

Seguidamente se presentan las estadísticas descriptivas del tercer periodo:

Ilustración 100

Estadísticas descriptivas del tercer periodo del reporte de Asignaturas Perdidas

```
numeric_perdidas[2].describe()
```

	NOTA_FINAL	VECES	CRED
count	3069.000000	3069.000000	3069.000000
mean	18.804822	1.212121	3.682796
std	10.500979	0.509655	2.125998
min	0.000000	1.000000	0.000000
25%	11.000000	1.000000	3.000000
50%	22.000000	1.000000	3.000000
75%	27.000000	1.000000	5.000000
max	47.000000	4.000000	20.000000

Las observaciones destacables de las estadísticas descriptivas de este periodo son las siguientes:

- Existen 3069 registros totales por cada variable numérica, lo que podría indicar que no existen valores nulos. Además, se puede observar que hubo más estudiantes reprobados en este semestre en comparación con los dos anteriores, específicamente 965 estudiantes más reprobados que en el periodo anterior.
- El promedio de notas finales de las asignaturas repetidas es de 18.80, lo que resulta relativamente bajo pero superior por 0.28 al periodo anterior. En general, se puede observar que el promedio de los estudiantes tiene un aumento de aproximadamente 0.31 puntos cada periodo.

- En promedio, las asignaturas que se perdieron son repetidas 1.2 veces, igual que en los periodos anteriores.
- En promedio, las asignaturas perdidas tienen un valor de 3.68 créditos. En general, este valor es el promedio entre los - periodos revisados hasta este momento.
- Dada la desviación estándar, es observable que la mayor parte de las notas se encontrarán entre 8.30 y 29.31.
- Dada la desviación estándar, se puede inferir que la mayor parte de registros referentes a la cantidad de veces que se repite una asignatura se da entre 0.70 y 1.73.
- Dada la desviación estándar, es factible inferir que la mayor parte de registros referentes a los créditos que valen cada asignatura se encuentran alrededor de 1.56 y 5.81.
- Se puede observar que en principio parecen no existir valores atípicos en los registros de las notas finales, teniendo como nota mínima 0 y máxima 47.
- Dada la observación de la cantidad mínima de repeticiones de una asignatura, se pueden dar dos suposiciones:
 - Se registra como repetición una vez que se pierde la asignatura.
 - Todas las asignaturas que se perdieron en este periodo son repetidas.
- Se puede observar que el 75% de las notas se encuentran por debajo o igual a 27, lo que podría significar que la nota máxima encontrada de 47 podría ser un valor atípico.

- Igualmente, se aplica para las repeticiones debido a que el 75% de sus datos son iguales a 1 y después, se tiene un valor máximo de 4 repeticiones, lo que resulta como más resaltante la posibilidad de que este valor sea un valor atípico debido a que, dadas las políticas de la universidad, no se puede repetir una sola asignatura más de 3 veces; sin embargo, dadas las observaciones hechas para la cantidad mínima de repeticiones, podría no considerarse como un valor atípico, sino como un referente de que el estudiante reprobó su tercera matrícula.
- Finalmente, es posible que igualmente puedan existir valores atípicos en el registro de créditos debido a que el 75% de sus registros poseen valores menores o iguales a 5 y se tiene un valor máximo de 20.

Finalmente, se presentan las estadísticas descriptivas del cuarto periodo:

Ilustración 101

Estadísticas descriptivas del cuarto periodo del reporte de Asignaturas Perdidas

```
numeric_perdidas[3].describe()
```

	NOTA_FINAL	VECES	CRED
count	3141.000000	3141.000000	3141.000000
mean	18.800328	1.207259	3.822429
std	10.124638	0.467435	2.152933
min	0.000000	1.000000	0.000000
25%	11.000000	1.000000	3.000000
50%	22.000000	1.000000	3.000000
75%	27.000000	1.000000	5.000000
max	49.000000	4.000000	20.000000

Las observaciones destacables de las estadísticas descriptivas de este periodo son las siguientes:

- Existen 3141 registros totales por cada variable numérica, lo que podría indicar que no existen valores nulos. Además, se puede observar que hubo más estudiantes reprobados en este semestre en comparación con los tres anteriores, pero fue el menor incremento con relación a los incrementos anteriores, específicamente hubo 72 estudiantes más reprobados que en el periodo anterior.
- El promedio de notas finales de las asignaturas repetidas se mantuvo en 18.80, lo que resulta relativamente bajo. En general, el patrón que se observaba de incremento en el promedio de notas finales no se cumplió para este promedio.
- En promedio, las asignaturas que se perdieron son repetidas 1.2 veces, igual que en los periodos anteriores.
- En promedio, las asignaturas perdidas tienen un valor de 3.82 créditos. En general, este valor es el promedio entre los periodos revisados hasta este momento.
- Dada la desviación estándar, es factible el considerar que la mayor parte de las notas se encontrarán entre 8.68 y 28.93.
- Dada la desviación estándar, es notable que existe la certitud de que la mayor parte de registros referentes a la cantidad de veces que se repite una asignatura se da entre 0.74 y 1.67.
- Dada la desviación estándar, se puede inferir que la mayor parte de registros referentes a los créditos que valen cada asignatura se encuentran alrededor de 1.67 y 5.98.

- Se puede observar que en principio parecen no existir valores atípicos en los registros de las notas finales, teniendo como nota mínima 0 y máxima 49.
- Dada la observación de la cantidad mínima de repeticiones de una asignatura, se pueden dar dos suposiciones:
 - Se registra como repetición una vez que se pierde la asignatura.
 - Todas las asignaturas que se perdieron en este periodo son repetidas.
- Se puede observar que el 75% de las notas se encuentran por debajo o igual a 27, lo que podría significar que la nota máxima encontrada de 49 podría ser un valor atípico.
- Igualmente, se aplica para las repeticiones debido a que el 75% de sus datos son iguales a 1 y después, se tiene un valor máximo de 4 repeticiones, lo que resulta como más resaltante la posibilidad de que este valor sea un valor atípico debido a que, dadas las políticas de la universidad, no se puede repetir una sola asignatura más de 3 veces; sin embargo, dadas las observaciones hechas para la cantidad mínima de repeticiones, podría no considerarse como un valor atípico, sino como un referente de que el estudiante reprobó su tercera matrícula.
- Finalmente, es posible observar que igualmente pueden existir valores atípicos en el registro de créditos debido a que el 75% de sus registros poseen valores menores o iguales a 5 y se tiene un valor máximo de 20.

A partir de todas las observaciones realizadas de todas las estadísticas descriptivas de todos los periodos del reporte de asignaturas perdidas, se lograron definir las siguientes conclusiones:

- En todos los periodos, se observa que es posible que no existan valores nulos en los registros, lo cual es positivo en términos de calidad de datos.
- El número de estudiantes que reprueban ha ido en aumento en cada periodo. El incremento más notorio se dio entre el periodo 2021-02 y 2022-01, con 965 estudiantes adicionales reprobados. El incremento más bajo se dio entre el periodo 2022-01 y 2022-02, con solo 72 estudiantes adicionales reprobados.
- El promedio de notas finales de las asignaturas repetidas se mantiene relativamente bajo en todos los periodos, oscilando entre 18.17 y 18.80. Se observó un incremento de alrededor de 0.31 puntos en cada periodo hasta el periodo 2022-01, pero este patrón no se mantuvo en el periodo 2022-02, donde el promedio se estancó en 18.80.
- En todos los periodos, las asignaturas perdidas se repiten en promedio 1.2 veces, es decir, es más usual que los estudiantes pierdan solo una vez una asignatura.
- El promedio de créditos de las asignaturas perdidas varía ligeramente entre los periodos, con valores que oscilan entre 3.68 y 3.82.
- En todos los periodos, se observan posibles valores atípicos en las notas finales, las repeticiones de asignaturas y los créditos de las asignaturas. Estos valores atípicos deben ser analizados con mayor detalle para determinar su impacto en los análisis y conclusiones.
- Las suposiciones relacionadas con las repeticiones de asignaturas (si se registra como repetición una vez que se pierde la asignatura o si todas las asignaturas que se perdieron en un periodo son repetidas) requieren una validación adicional para asegurar una interpretación adecuada de los datos.

Hasta este momento se ha realizado el primer análisis explorando las variables numéricas del reporte de asignaturas perdidas, enseguida se presentarán el código, las ilustraciones y observaciones de los histogramas realizados a las variables numéricas del reporte, cabe resaltar que, como se ha mencionado previamente, los histogramas permiten identificar la distribución, modas, posibles sesgos y valores atípicos en los datos, lo que facilita la detección de patrones y comportamientos particulares en cada variable.

Se iniciará con la visualización de los histogramas de las notas finales:

Ilustración 102

Código de la generación de un histograma de las notas finales del reporte de Asignaturas Perdidas

```
# Se crea una figura y ejes
fig, axes = plt.subplots(2, 2, figsize=(12, 10))

periodos = ["2021-01", "2021-02", "2022-01", "2022-02"]

# Se itera sobre los índices de las subfiguras y los datos correspondientes
for i, ax in enumerate(axes.flatten()):
    # Se crea una tabla de frecuencia para el dataset actual
    tabla_frecuencia = pd.cut(numeric_perdidas[i]['NOTA_FINAL'], bins=np.arange(0, 55, 5)).value_counts(sort=False)

    # Se crea un histograma a partir de la tabla de frecuencia y se dibuja en la subfigura actual
    ax.bar(x=tabla_frecuencia.index.astype(str), height=tabla_frecuencia.values)

    # Se agregan etiquetas y título a la subfigura actual
    ax.set_xlabel("Rangos de notas finales")
    ax.set_ylabel("Frecuencia")
    ax.set_title(f"Periodo {periodos[i]}")

    # Se giran verticalmente los xticks
    ax.tick_params(axis='x', rotation=90)

# Se ajusta el espacio entre las subfiguras
plt.tight_layout()

# Se muestra el gráfico completo que contiene las subfiguras
plt.show()
```

Ilustración 103

Histogramas del primer y segundo periodo de la variable NOTA_FINAL del reporte de asignaturas perdidas

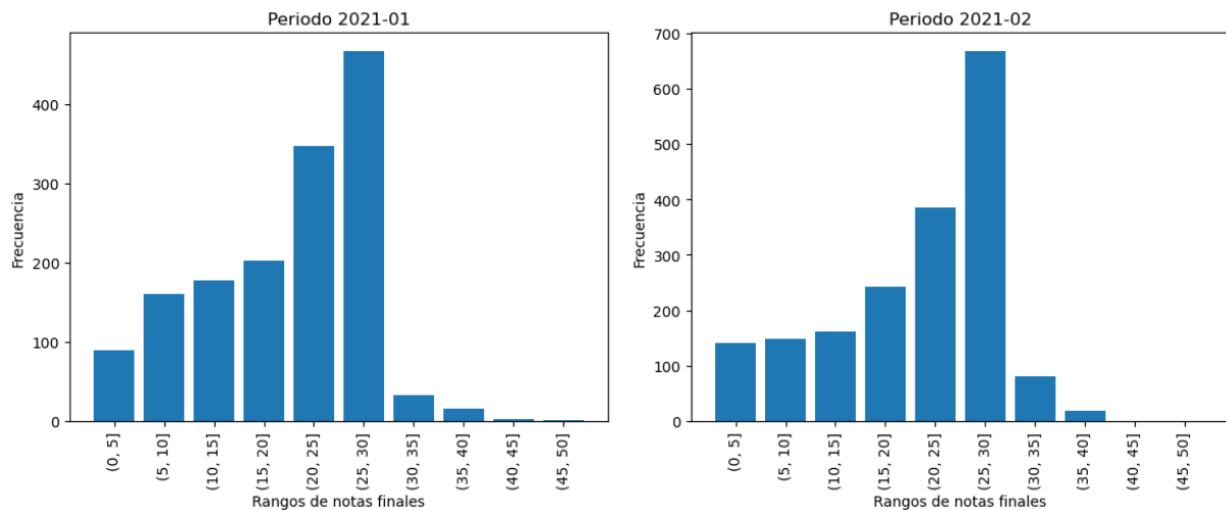
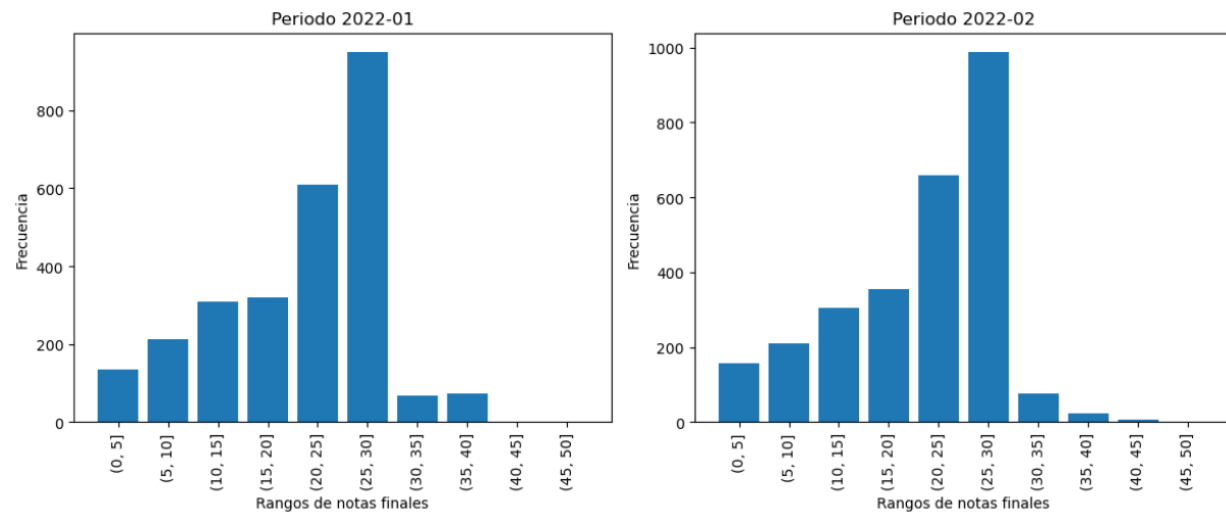


Ilustración 104

Histogramas del tercer y cuarto periodo de la variable NOTA_FINAL del reporte de asignaturas perdidas



A partir de los histogramas visualizados se pudieron determinar las siguientes conclusiones:

- Todos los períodos muestran una distribución similar de calificaciones finales, de hecho, en todos los períodos se cumple que las calificaciones finales más frecuentes oscilan entre 25 y 30.
- Se puede observar que existen estudiantes con calificaciones finales superiores a 30, por lo que sugiere que estos estudiantes no perdieron sus asignaturas únicamente por su desempeño en términos de calificaciones, sino que otros factores, como faltas, podrían haber influido.
- El gráfico de distribución permite visualizar claramente que las calificaciones superiores a 30 son mucho menos comunes, mientras que la gran mayoría corresponde a calificaciones que podrían considerarse normales.

Se proseguirá con la visualización de los histogramas de la cantidad de veces que un estudiante repitió una misma asignatura:

Ilustración 105

Código de la generación de un histograma de las repeticiones de un estudiante en una determinada asignatura del reporte de Asignaturas Perdidas

```
# Se crea una figura y ejes
fig, axes = plt.subplots(2, 2, figsize=(12, 10))

periodos = ["2021-01", "2021-02", "2022-01", "2022-02"]

# Se itera sobre los índices de las subfiguras y los datos correspondientes
for i, ax in enumerate(axes.flatten()):
    # Se crea una tabla de frecuencia para el dataset actual
    tabla_frecuencia = pd.cut(numeric_perdidas[i]['VECES'], bins=np.arange(0, 6, 1)).value_counts(sort=False)

    # Se crea un histograma a partir de la tabla de frecuencia y se dibuja en la subfigura actual
    ax.bar(x=tabla_frecuencia.index.astype(str), height=tabla_frecuencia.values)

    # Se agregan etiquetas y título a la subfigura actual
    ax.set_xlabel("Rangos de cantidad de veces que un estudiante ha repetido una asignatura")
    ax.set_ylabel("Frecuencia")
    ax.set_title(f"Periodo {periodos[i]}")

    # Se giran verticalmente los xticks
    ax.tick_params(axis='x', rotation=90)

# Se ajusta el espacio entre las subfiguras
plt.tight_layout()

# Se muestra el gráfico completo que contiene las subfiguras
plt.show()
```

Ilustración 106

Histogramas del primer y segundo periodo de la variable VECES del reporte de Asignaturas Perdidas

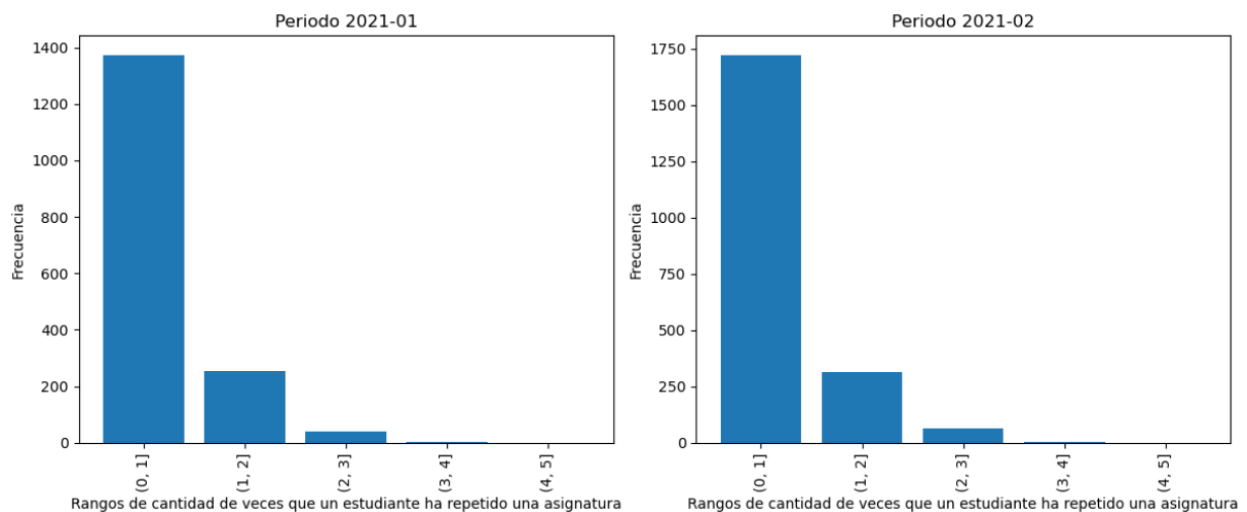
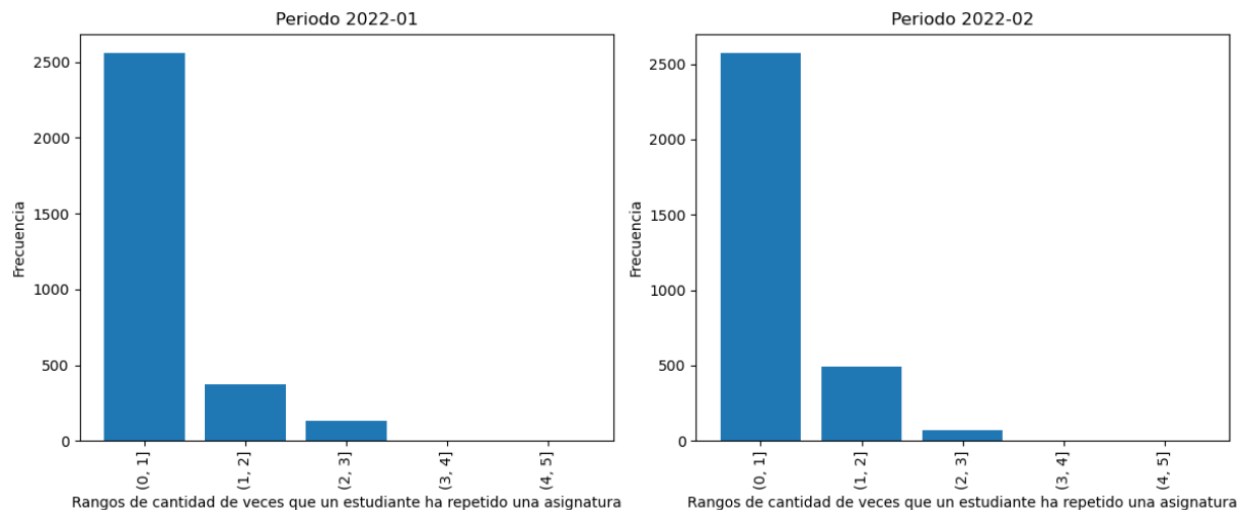


Ilustración 107

Histogramas del tercer y cuarto periodo de la variable VECES del reporte de Asignaturas Perdidas



A partir de los histogramas visualizados se pudieron determinar las siguientes conclusiones:

- Se puede observar de mejor manera cómo en todos los períodos se cumple que la mayor cantidad de los registros representan a estudiantes reprobados por primera vez.
- Dada la distribución de datos, se puede inferir lo siguiente: cuando un estudiante reprueba una asignatura, a este se le suma 1 a la cantidad actual de repeticiones en la asignatura dada. Es decir, si un estudiante que toma por primera vez una asignatura reprueba, a este se le asigna $1 + 0$ (puesto que toma por primera vez la asignatura y nunca la ha reprobado) al reporte de asignaturas perdidas; por otra parte, en el caso de que el mismo estudiante vuelva a reprobear esta asignatura en el siguiente período, a este se le asignaría $1 + 1$ (puesto que se encontraba en su primera repetición) al reporte de asignaturas perdidas de aquel período. De esta manera, nunca se poseen valores de 0 ni de 5 en el reporte, pero existe la rara posibilidad de un 4, puesto que, si un

estudiante reprueba su tercera matrícula en una asignatura, se le asignaría en el reporte un total de 1 + 3 (debido a que se encontraba en su tercera repetición).

- Es evidente que existe una menor probabilidad de que un estudiante repruebe una asignatura que ya ha reprobado anteriormente. Por ejemplo, si un estudiante reprueba una vez alguna asignatura, es menos probable que la repruebe dos veces. Esta lógica se aplica a medida que aumentan las repeticiones de una asignatura por parte de un estudiante.

Finalmente, se presentan los histogramas de la cantidad de créditos por asignatura:

Ilustración 108

Código de la generación de un histograma de los créditos por asignatura del reporte de Asignaturas Perdidas

```
# Se crea una figura y ejes
fig, axes = plt.subplots(4, 1, figsize=(12, 30))

periodos = ["2021-01", "2021-02", "2022-01", "2022-02"]

# Se itera sobre los índices de las subfiguras y los datos correspondientes
for i, ax in enumerate(axes.flatten()):
    # Se crea una tabla de frecuencia para el dataset actual
    tabla_frecuencia = pd.cut(numeric_perdidas[i]['CRED'], bins=np.arange(0, 21, 1)).value_counts(sort=False)

    # Se crea un histograma a partir de la tabla de frecuencia y se dibuja en la subfigura actual
    ax.bar(x=tabla_frecuencia.index.astype(str), height=tabla_frecuencia.values)

    # Se agregan etiquetas y título a la subfigura actual
    ax.set_xlabel("Rangos de créditos de asignaturas")
    ax.set_ylabel("Frecuencia")
    ax.set_title(f"Periodo {periodos[i]}")

    # Se giran verticalmente los xticks
    ax.tick_params(axis='x', rotation=90)

# Se ajusta el espacio entre las subfiguras
plt.tight_layout()

# Se muestra el gráfico completo que contiene las subfiguras
plt.show()
```

Ilustración 109

Histograma del primer periodo de la variable CRED del reporte de Asignaturas Perdidas

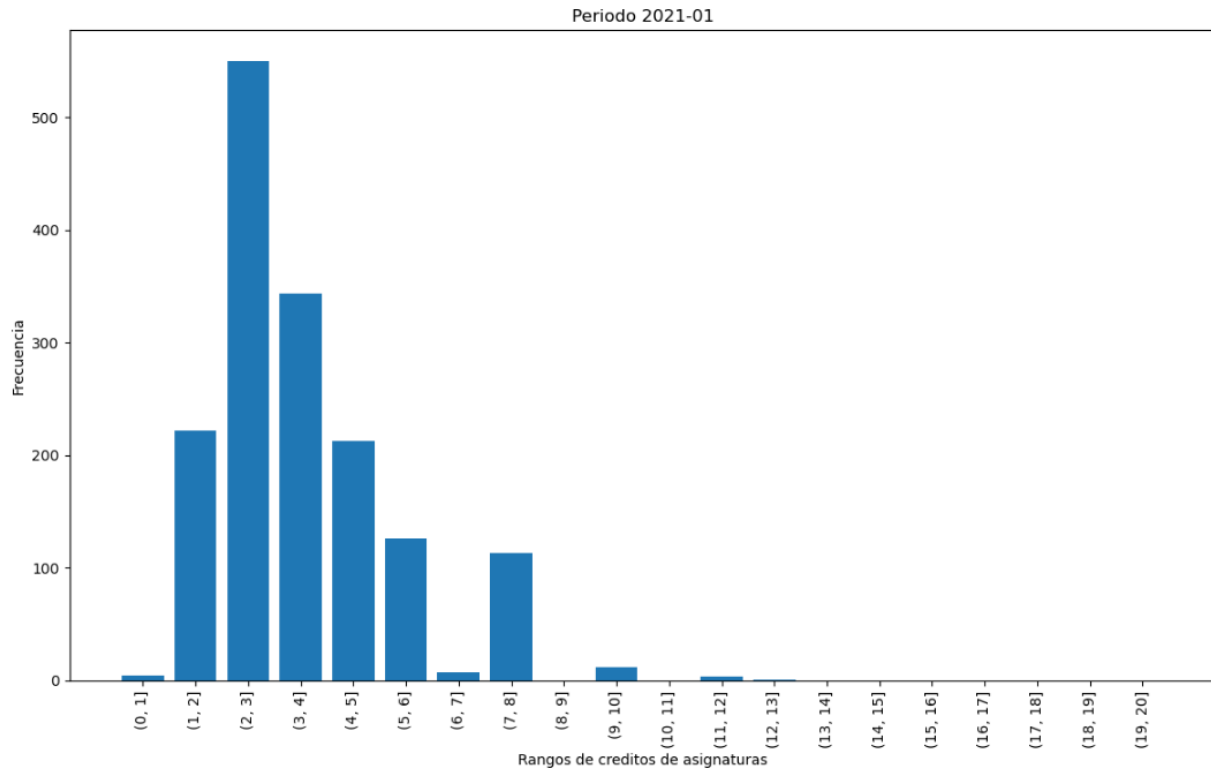


Ilustración 110

Histograma del segundo periodo de la variable CRED del reporte de Asignaturas Perdidas

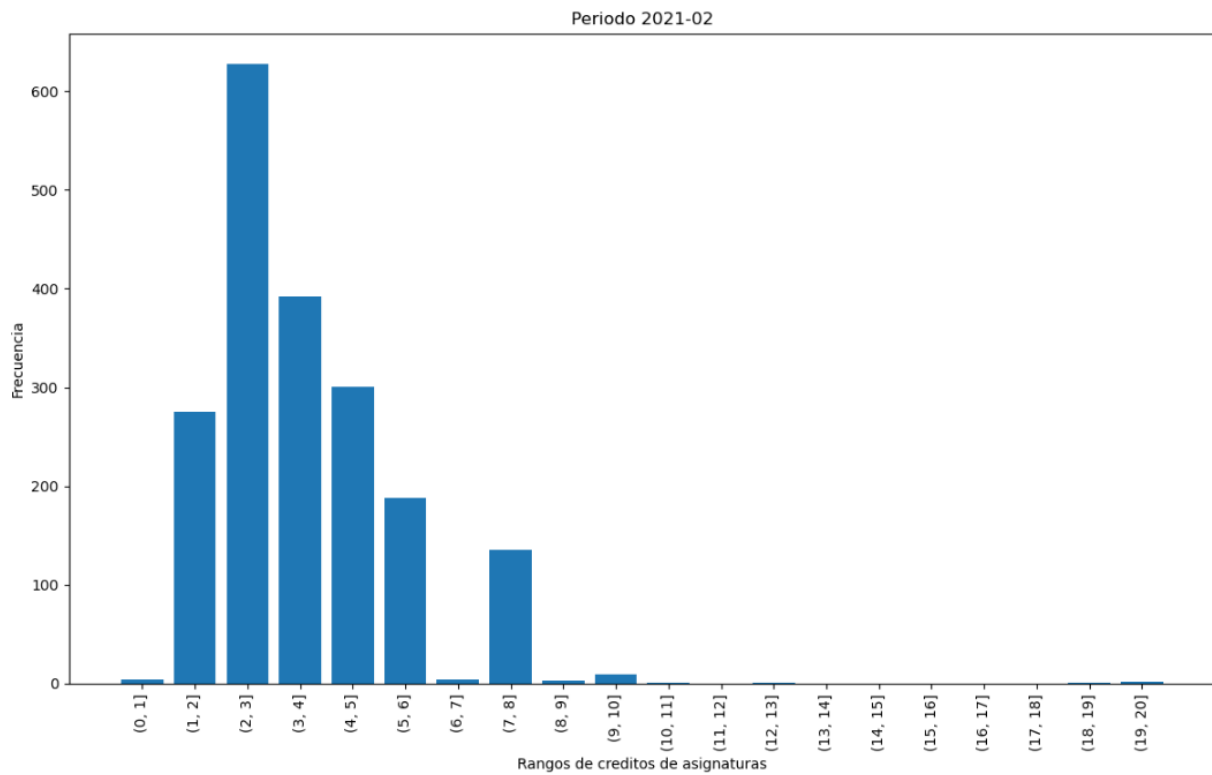


Ilustración 111

Histograma del tercer periodo de la variable CRED del reporte de Asignaturas Perdidas

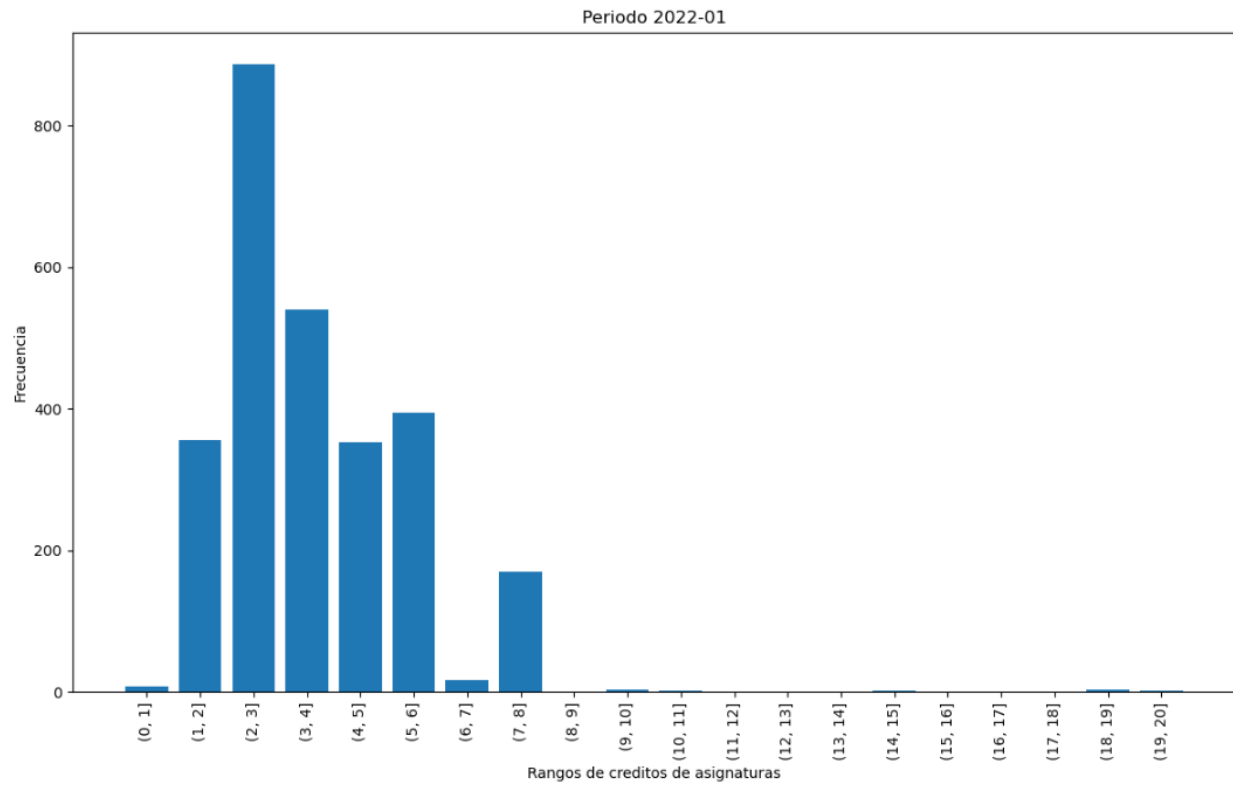
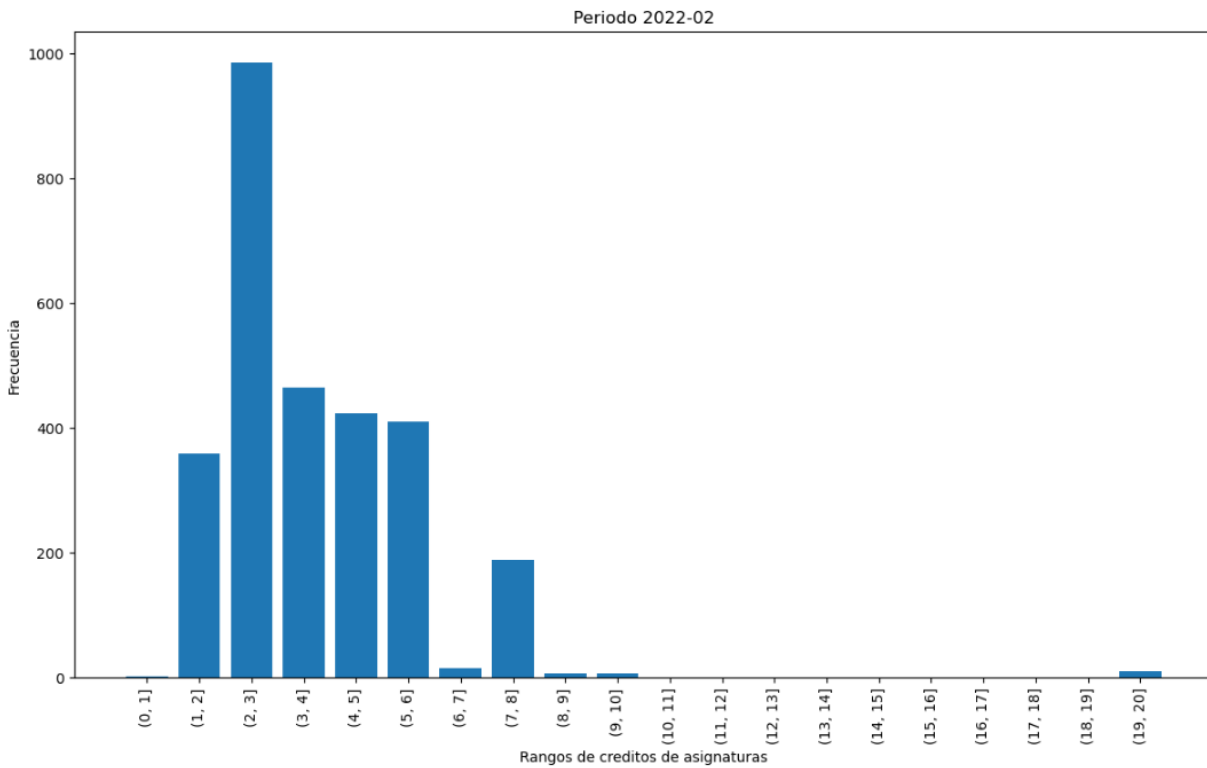


Ilustración 112

Histograma del cuarto periodo de la variable CRED del reporte de Asignaturas Perdidas



A partir de los histogramas visualizados se pudieron determinar las siguientes conclusiones:

- Se puede observar que la mayor cantidad de asignaturas reprobadas poseen 3 créditos, y existe una distribución similar para las asignaturas con 2, 4, 5 y 6 créditos en todos los períodos.
- Se puede observar que la distribución de asignaturas reprobadas con un total de 8 no muestra una variación significativa entre períodos, lo que podría indicar una relación fuerte.

- Se puede observar que en todos los períodos existen posibles valores atípicos. Esto puede significar que hay pocas asignaturas con esta cantidad de créditos o simplemente tienen una menor probabilidad de ser reprobadas por los estudiantes.

Consecuentemente, una vez finalizado el proceso de exploración de datos por histogramas, se continuará con la exploración haciendo uso de gráficos de densidad, que servirá para presentar una visión más detallada de la distribución de los datos en comparación a los histogramas presentados.

Para empezar, se visualizará el gráfico de densidad de las finales del reporte de asignaturas perdidas:

Ilustración 113

Código que genera el gráfico de densidad de las notas finales del reporte de Asignaturas Perdidas

```
# Generacion de un dataframe con datos por periodo academico
datos_temp = pd.DataFrame()
datos_temp["PERIODO 2021-01"] = numeric_perdidas[0]['NOTA_FINAL']
datos_temp["PERIODO 2021-02"] = numeric_perdidas[1]['NOTA_FINAL']
datos_temp["PERIODO 2022-01"] = numeric_perdidas[2]['NOTA_FINAL']
datos_temp["PERIODO 2022-02"] = numeric_perdidas[3]['NOTA_FINAL']

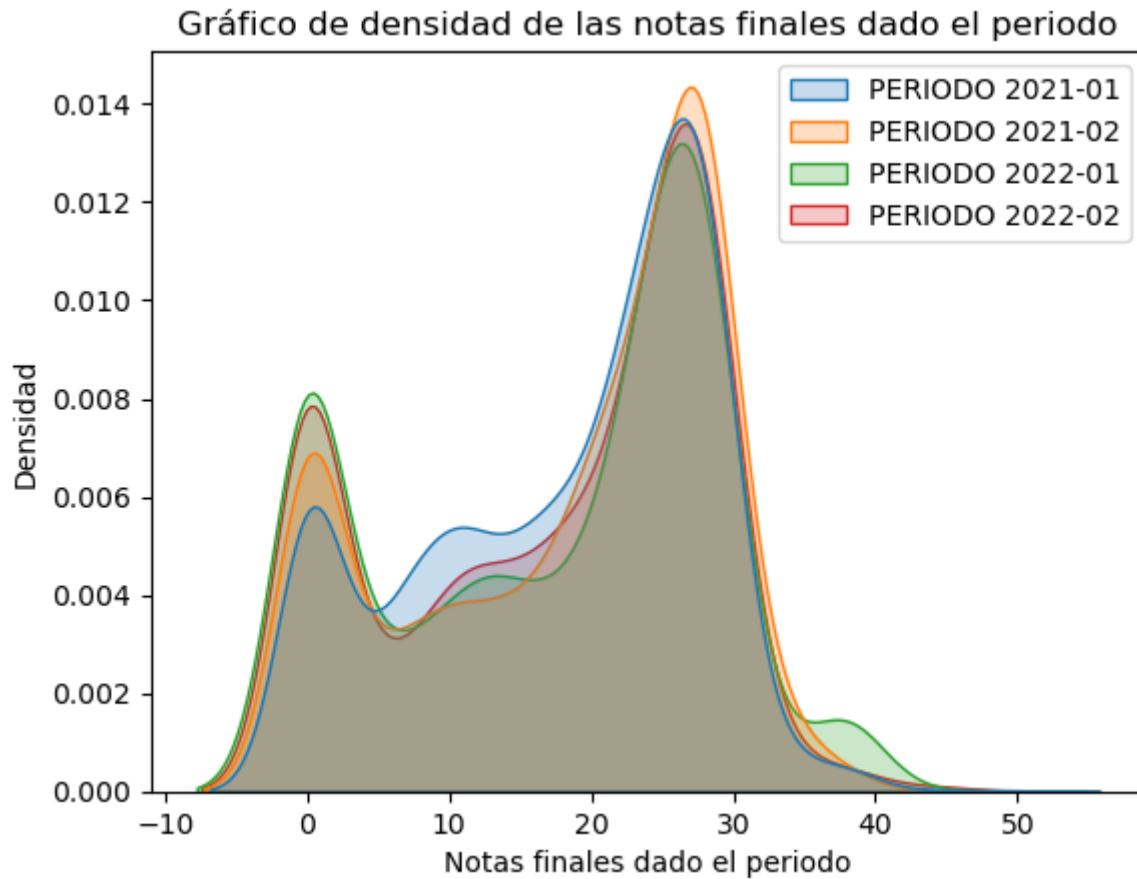
# Crear un gráfico de densidad utilizando Seaborn
sns.kdeplot(data=datos_temp, shade=True)

# Agregar etiquetas y título
plt.xlabel('Notas finales dado el periodo')
plt.ylabel('Densidad')
plt.title('Gráfico de densidad de las notas finales dado el periodo')

# Mostrar el gráfico
plt.show()
```

Ilustración 114

Gráfico de densidad de la variable NOTA_FINAL del reporte de Asignaturas Finales



A partir de los gráficos de densidad visualizados se pudieron determinar las siguientes conclusiones:

- El periodo 2022-01 posee una mayor cantidad de calificaciones finales con 0, mientras que el periodo 2021-01 tuvo la menor cantidad de calificaciones con 0, lo que indica que hubo un posible descenso en el rendimiento académico general de los estudiantes a través de los periodos académicos.

- Se pueden observar varios valores atípicos, tanto inferiores a 0 como superiores a 50. Sin embargo, estos valores son únicamente el resultado de la representación gráfica y no corresponden a valores reales.
- Se puede observar cómo a lo largo de los periodos, la distribución de calificaciones ha tendido hacia mejores resultados, sin embargo, la comparación entre periodos no sugiere lo mismo pues el primer periodo que se estuvo completamente en modalidad virtual posee la mayor cantidad de mejores calificaciones de asignaturas perdidas.
- Se puede visualizar una clara distribución de las notas finales en valores entre 20 y 30, lo cual coincide con el análisis realizado mediante los histogramas.

Se seguirá con a la visualización del gráfico de densidad de la cantidad de veces que un estudiante repite la misma asignatura:

Ilustración 115

Código que genera el gráfico de densidad de la cantidad de veces que un estudiante repitió una misma asignatura del reporte de Asignaturas Perdidas

```
# Generacion de un dataframe con datos por periodo academico
datos_temp = pd.DataFrame()
datos_temp["PERIODO 2021-01"] = numeric_perdidas[0]['VECES']
datos_temp["PERIODO 2021-02"] = numeric_perdidas[1]['VECES']
datos_temp["PERIODO 2022-01"] = numeric_perdidas[2]['VECES']
datos_temp["PERIODO 2022-02"] = numeric_perdidas[3]['VECES']

# Crear un gráfico de densidad utilizando Seaborn
sns.kdeplot(data=datos_temp, shade=True)

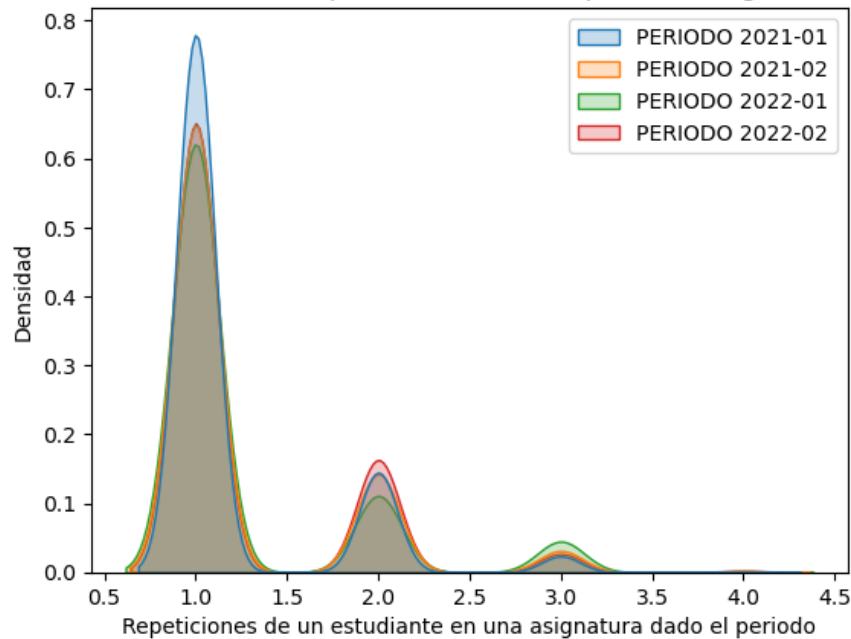
# Agregar etiquetas y título
plt.xlabel('Repeticiones de un estudiante en una asignatura dado el periodo')
plt.ylabel('Densidad')
plt.title('Gráfico de densidad de las veces que un estudiante repitió una asignatura dado el periodo')

# Mostrar el gráfico
plt.show()
```

Ilustración 116

Gráfico de densidad de la variable VECES del reporte de Asignaturas Perdidas

Gráfico de densidad de las veces que un estudiante repitió una asignatura dado el periodo



A partir del gráfico de densidad, se pudieron determinar la siguiente conclusión:

- El periodo 2021-01 tiene la mayor cantidad de estudiantes que han repetido por primera vez una asignatura. Como consecuencia de esto, se puede apreciar que, en el siguiente periodo, solo un pequeño porcentaje de los estudiantes que reprobaron por primera vez volvieron a reprobar. Este patrón se mantiene a medida que aumenta el número de veces que se reprueba la asignatura.

Finalmente, se presenta el gráfico de densidad de la cantidad créditos que posee cada asignatura:

Ilustración 117

Código que genera el gráfico de densidad de la cantidad de créditos de asignaturas del reporte de Asignaturas Perdidas

```
# Generacion de un dataframe con datos por periodo academico
datos_temp = pd.DataFrame()
datos_temp["PERIODO 2021-01"] = numeric_perdidas[0]['CRED']
datos_temp["PERIODO 2021-02"] = numeric_perdidas[1]['CRED']
datos_temp["PERIODO 2022-01"] = numeric_perdidas[2]['CRED']
datos_temp["PERIODO 2022-02"] = numeric_perdidas[3]['CRED']

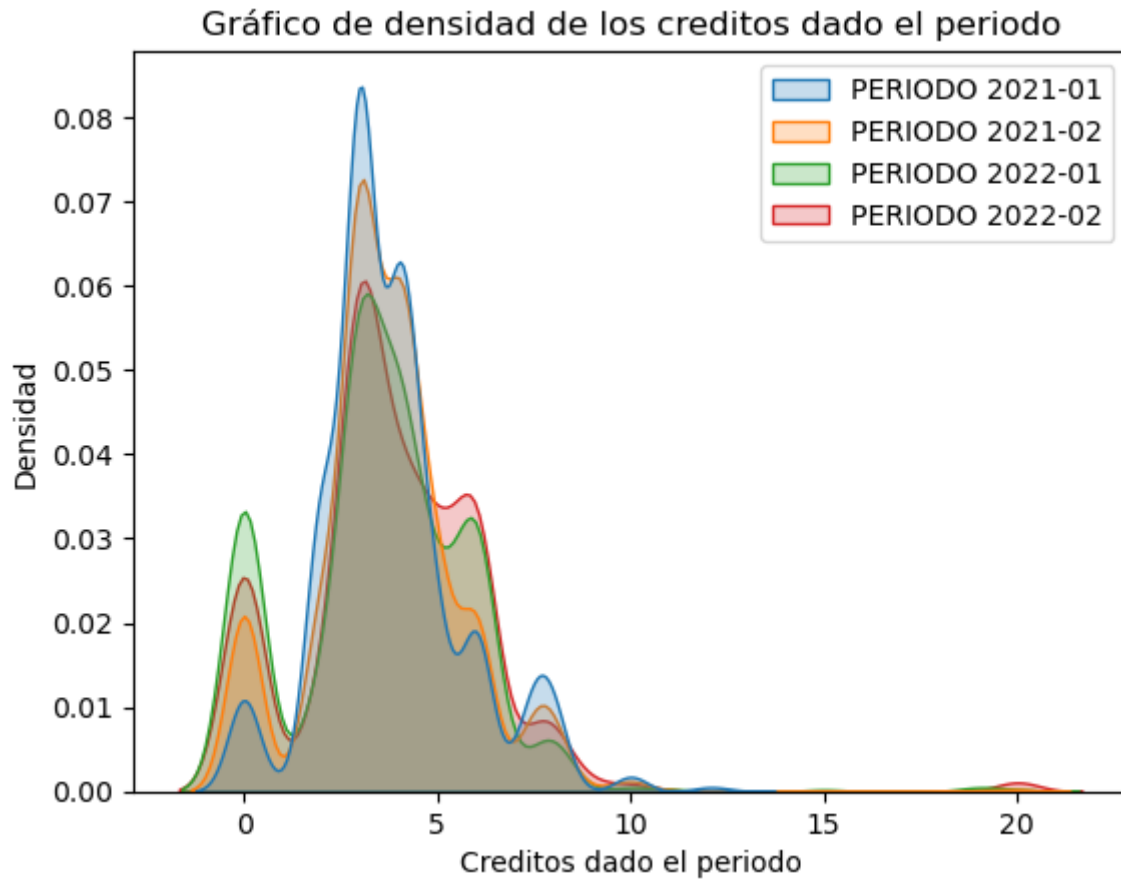
# Crear un gráfico de densidad utilizando Seaborn
sns.kdeplot(data=datos_temp, shade=True)

# Agregar etiquetas y título
plt.xlabel('Creditos dado el periodo')
plt.ylabel('Densidad')
plt.title('Gráfico de densidad de los creditos dado el periodo')

# Mostrar el gráfico
plt.show()
```


Ilustración 118

Gráfico de densidad de la variable CRED del reporte de Asignaturas Perdidas



Dado el gráfico de densidad de la cantidad de créditos que poseen las asignaturas reprobadas a través del tiempo, se logró determinar la siguiente conclusión:

- Se puede concluir que hay una distinción clara entre las asignaturas reprobadas y su cantidad de créditos entre los periodos, ya que varía considerablemente en la mayoría de los casos. Sin embargo, se puede notar una disminución en la variabilidad de las asignaturas con un valor cercano a 3 créditos. A pesar de esto, el análisis sugiere que posiblemente no exista una relación definitiva entre la cantidad de créditos y las calificaciones finales, aunque no se puede llegar a una conclusión concluyente.

Por consiguiente, a los análisis exploratorios realizados a este momento, ha llegado el momento de realizar los diagramas de Caja y Bigotes, sin embargo, es necesario comentar que un diagrama de este tipo no resulta útil para variables discreta, por lo que los diagramas se realizarán sobre variables continuas, es decir, NOTA_FINAL. Este gráfico podrá ser apreciado a continuación con sus respectivas conclusiones:

Ilustración 119

Código para la generación de diagramas de caja y bigotes de las notas finales del reporte de Asignaturas Perdidas

```
plt.figure(figsize=(10, 8))

plt.subplot(221)
sns.boxplot(y='NOTA_FINAL', data=numeric_perdidas[0])
plt.title("Diagrama de caja y bigotes de las notas finales - 2021-01")
plt.ylabel('Nota final')

plt.subplot(222)
sns.boxplot(y='NOTA_FINAL', data=numeric_perdidas[1])
plt.title("Diagrama de caja y bigotes de las notas finales - 2021-02")
plt.ylabel('Nota final')

plt.subplot(223)
sns.boxplot(y='NOTA_FINAL', data=numeric_perdidas[2])
plt.title("Diagrama de caja y bigotes de las notas finales - 2022-01")
plt.ylabel('Nota final')

plt.subplot(224)
sns.boxplot(y='NOTA_FINAL', data=numeric_perdidas[3])
plt.title("Diagrama de caja y bigotes de las notas finales - 2022-02")
plt.ylabel('Nota final')

plt.tight_layout()
plt.show()
```

Ilustración 120

Diagramas de caja y bigotes del primer y segundo periodo de la variable NOTA_FINAL del reporte de Asignaturas Perdidas

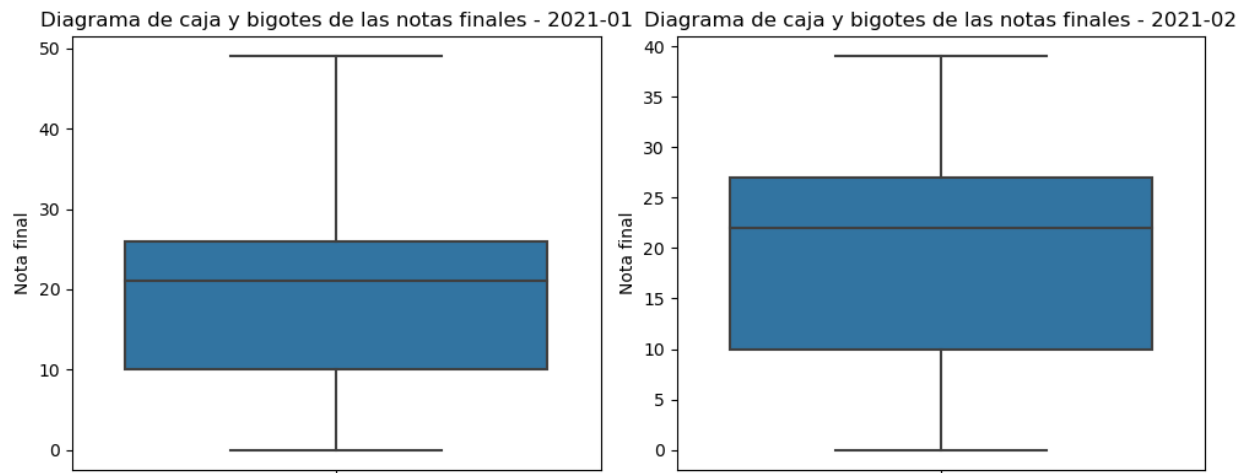
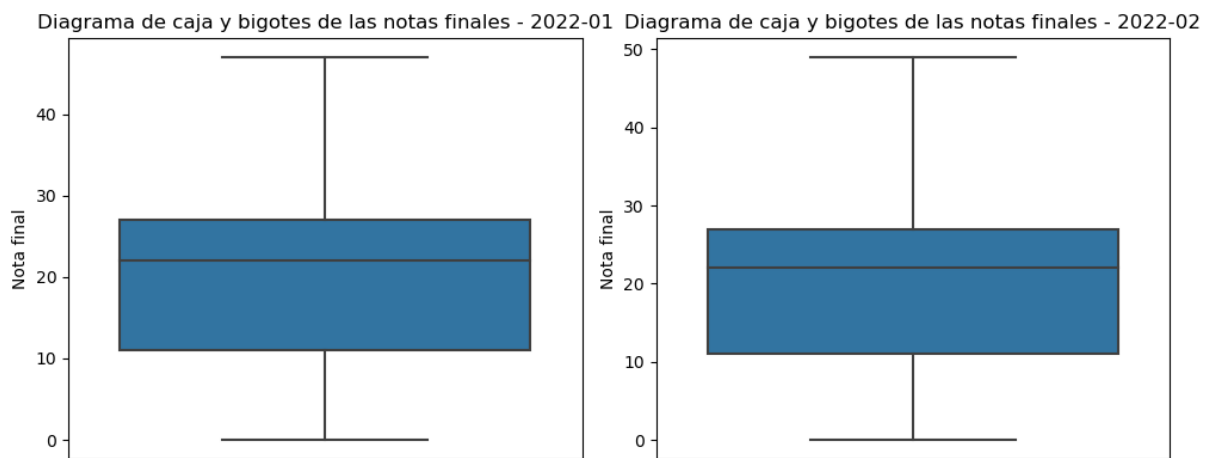


Ilustración 121

Diagramas de caja y bigotes del tercer y cuarto periodo de la variable NOT_FINAL del reporte de Asignaturas Perdidas



Dado los diagramas de caja y bigotes presentados, se pudo determinar las siguientes conclusiones:

- Se puede observar como la distribución de las calificaciones es sumamente semejante a través de los periodos académicos, lo que sugiere una alta posibilidad de la existencia de una relación entre una materia reprobada y su nota final. (Nota del autor:

Resulta un poco obvia esta conclusión ya que el resultado de la calificación final influye directamente en si se aprobó o reprobó una asignatura, sin embargo, es importante notar que no es la única razón por la que un estudiante puede reprobear una asignatura como lo sería el ausentismo.)

- Es gratificante notar que hay pocos estudiantes con una calificación final de 0, lo que indica que, a pesar de las dificultades académicas, la mayoría de los estudiantes se esfuerzan por mejorar.
- Se podría inferir que algunos estudiantes tienen un ausentismo continuo, lo cual resulta en su reprobación. Lamentablemente, en los datos compartidos para esta investigación, no se cuentan con registros de estas faltas, lo que plantea la necesidad de tomar una decisión sobre qué hacer con estos datos. Se debe considerar si se asume que todas las reprobaciones con calificaciones mayor a 30 son resultado de faltas o si se eliminan estos casos.

Finalmente se ha finalizado la exploración de variables numéricas del reporte de asignaturas perdidas, por lo que es necesario mencionar que se logró la identificación de las siguientes variables de interés de este reporte con lo que respecta a variables numéricas:

- NOTA_FINAL
- CRED
- VECES

Esta selección de variables de interés indica que estas variables podrían ser consideradas para el modelo, sin embargo, hace falta un mayor análisis, en específico una exploración de su relación con el rendimiento académico.

A partir de la finalización del análisis exploratorio de las variables numéricas del reporte de asignaturas perdidas, es momento idóneo para continuar la exploración de variables con el análisis del reporte de calificaciones por carrera comenzando por sus estadísticas descriptivas.

Para empezar, a continuación, se despliegan las estadísticas descriptivas del primer periodo del reporte:

Ilustración 122

Estadísticas descriptivas del primer periodo del reporte de Calificaciones por Carrera

```
numeric_calificaciones[0].describe()
```

	nota_parcial	nota_final
count	163771.000000	163806.000000
mean	41.711148	41.713082
std	8.389080	6.918972
min	0.000000	0.000000
25%	38.050000	38.600000
50%	43.800000	43.050000
75%	48.000000	46.490000
max	50.000000	50.000000

A partir de las estadísticas descriptivas demostradas, se pudo recopilar las siguientes observaciones:

- Se observa la posibilidad de que existan valores nulos, ya que hay 163,771 notas parciales, pero 163,806 notas finales.

- Tanto el promedio de las notas parciales como el de las notas finales es muy similar, con un valor de 41.71. Esto indica un alto nivel académico promedio en la universidad. Además, se aprecia una relación entre las notas parciales y las finales (lo cual es bastante obvio, ya que las notas parciales determinan la nota final).
- Basándose en la desviación estándar, se puede inferir que la mayoría de las notas parciales se encuentran entre 33.32 y 50.
- De igual manera, basándose en la desviación estándar, se puede inferir que la mayoría de las notas finales se encuentran entre 34.81 y 48.62.
- El 25% de las notas finales se sitúa por debajo de los 38.6 puntos, lo cual es bastante similar a las notas parciales, que se encuentran por debajo de los 30.05 puntos.
- En general, se puede observar que para el 50% de los datos, las calificaciones son inferiores a 43 puntos.
- El 75% de las notas parciales son inferiores a 48, mientras que las notas finales tienen un puntaje promedio de 46.49.
- Por último, no se observan valores atípicos, ya que los valores mínimos son 0 y los máximos son 50, lo cual está dentro del rango esperado.

A continuación, es momento de presentar las estadísticas descriptivas del segundo periodo del reporte de calificaciones por carrera:

Ilustración 123

Estadísticas descriptivas del segundo periodo del reporte de Calificaciones por Carrera

```
numeric_calificaciones[1].describe()
```

	nota_parcial	nota_final
count	158628.000000	158818.000000
mean	41.308094	41.302036
std	8.967292	7.499267
min	0.000000	0.000000
25%	37.500000	37.860000
50%	43.540000	42.735000
75%	48.000000	46.510000
max	50.000000	50.000000

A partir de las estadísticas descriptivas desplegadas, se pudo observar lo siguiente:

- Se observa la posibilidad de que existan valores nulos, ya que hay 158,628 notas parciales, pero 158,818 notas finales.
- Tanto el promedio de las notas parciales como el de las notas finales es muy similar, con un valor de 41.31 y 41.30 respectivamente. Esto indica un alto nivel académico promedio en la universidad. Sin embargo, también indica que el promedio con respecto al pasado ha decrecido.
- La desviación estándar de las notas parciales es 8.97, lo que implica que la mayoría de las notas parciales están en el rango de 32.34 a 50.00. La desviación estándar de las notas finales es de 7.50, lo que sugiere que la mayoría de las notas finales están en el rango de 33.80 a 48.80.

- El 25% de las notas finales se sitúa por debajo de los 37.86 puntos, lo cual es bastante similar a las notas parciales, que se encuentran por debajo de los 37.5 puntos.
- La mediana (el percentil 50) para las notas parciales es de 43.54 y para las notas finales es de 42.74. Esto significa que alrededor del 50% de los estudiantes tienen calificaciones por debajo de estos valores.
- El 75% de las notas parciales son inferiores a 48, mientras que las notas finales tienen un puntaje promedio de 46.5.
- Por último, no se observan valores atípicos, ya que los valores mínimos son 0 y los máximos son 50, lo cual está dentro del rango esperado.

Seguido de estas conclusiones, se continuará con la muestra de las estadísticas descriptivas del tercer periodo del reporte de calificaciones por carrera:

Ilustración 124

Estadísticas descriptivas del tercer periodo del reporte de Calificaciones por Carrera

```
numeric_calificaciones[2].describe()
```

	nota_parcial	nota_final
count	150107.000000	150108.000000
mean	39.622325	39.620595
std	10.135256	8.522749
min	0.000000	0.000000
25%	35.000000	35.730000
50%	42.000000	41.040000
75%	47.000000	45.610000
max	50.000000	50.000000

A partir de estas estadísticas descriptivas, se observaron los siguientes comportamientos:

- Existe la posibilidad de que haya un único valor nulo, ya que el conteo de notas parciales es de 150,107, mientras que el conteo de notas finales es de 150,108.
- El promedio de las notas parciales y las notas finales es muy similar, siendo 39.62 para ambos. Esto indica que el rendimiento académico promedio sigue siendo alto, aunque hay una más marcada disminución en comparación a los periodos anteriores, exactamente de 1.69 puntos a diferencia de la anterior disminución de a pesar 0.4 puntos.
- La desviación estándar de las notas parciales es 10.14, lo que implica que la mayoría de las notas parciales están en el rango de 29.49 a 50.00. La desviación estándar de las notas finales es de 8.52, lo que sugiere que la mayoría de las notas finales están en el rango de 31.10 a 48.14.
- El 25% de las notas finales se sitúa por debajo de los 35.73 puntos, mientras que las notas parciales están por debajo de los 35 puntos. Esto sugiere que la distribución de las notas ha cambiado ligeramente en comparación con el conjunto de datos anterior.
- La mediana de las notas parciales es de 42.00 y para las notas finales es de 41.04. Esto significa que alrededor del 50% de los estudiantes tienen calificaciones por debajo de estos valores.
- El 75% de las notas parciales son inferiores a 47, mientras que las notas finales tienen un puntaje promedio de 45.61. Esto muestra una pequeña disminución en comparación con el conjunto de datos anterior.

- Por último, no se observan valores atípicos, ya que los valores mínimos son 0 y los máximos son 50, lo cual está dentro del rango esperado.

Finalmente, es momento de presentar las estadísticas descriptivas que corresponden al último periodo del reporte de calificaciones por carrera:

Ilustración 125

Estadísticas descriptivas del cuarto periodo del reporte de Calificaciones por Carrera

```
numeric_calificaciones[3].describe()
```

	nota_parcial	nota_final
count	142200.000000	142189.000000
mean	38.918104	38.914116
std	10.348763	8.798959
min	0.000000	0.000000
25%	34.160000	34.870000
50%	41.000000	40.290000
75%	46.610000	45.000000
max	50.000000	50.000000

A partir de estas últimas estadísticas descriptivas, se logró observar lo siguiente:

- Existe la posibilidad de que haya 11 valores nulos, ya que el conteo de notas parciales es de 142,200, mientras que el conteo de notas finales es de 142,189.
- El promedio de las notas parciales y las notas finales es muy similar, siendo 38.92 y 38.91 respectivamente. Esto indica un rendimiento académico promedio moderado, con una disminución en comparación a los periodos anteriores, exactamente de 0.71 puntos.

- La desviación estándar de las notas parciales es 10.35, lo que implica que la mayoría de las notas parciales están en el rango de 28.57 a 50.00. La desviación estándar de las notas finales es de 8.80, lo que sugiere que la mayoría de las notas finales están en el rango de 30.12 a 47.71.
- El 25% de las notas finales se sitúa por debajo de los 34.87 puntos, mientras que las notas parciales están por debajo de los 34.16 puntos. Esto sugiere que la distribución de las notas ha cambiado ligeramente en comparación con el conjunto de datos anterior.
- La mediana de las notas parciales es de 41.00 y para las notas finales es de 40.29. Esto significa que alrededor del 50% de los estudiantes tienen calificaciones por debajo de estos valores.
- El 75% de las notas parciales son inferiores a 46.61, mientras que las notas finales tienen un puntaje promedio de 45.00. Esto muestra una pequeña disminución en comparación con el conjunto de datos anterior.
- Por último, no se observan valores atípicos, ya que los valores mínimos son 0 y los máximos son 50, lo cual está dentro del rango esperado para las notas de los estudiantes.

Consecuente a la finalización del análisis de las estadísticas descriptivas de cada periodo del reporte de calificaciones por carreras, se pudieron determinar las siguientes conclusiones:

- Hay evidencia de la existencia de posibles valores nulos en todos los periodos observados, lo que podría indicar la presencia de datos perdidos o estudiantes que no completaron sus calificaciones.

- Durante los periodos observados, se ha producido una disminución constante en el promedio de las notas tanto parciales como finales. Esto podría indicar un declive general en el rendimiento académico, o cambios en la estructura del currículo o los estándares de calificación.
- La desviación estándar de las notas parciales y finales ha aumentado en cada periodo, lo que indica un mayor grado de variabilidad en las calificaciones de los estudiantes. Esto podría sugerir una creciente divergencia en el rendimiento académico de los estudiantes.
- La mediana y los percentiles (25%, 50%, y 75%) de las notas parciales y finales también han mostrado una tendencia decreciente, lo que indica que la distribución general de las calificaciones se está moviendo hacia el extremo inferior de la escala. Esto podría reflejar un rendimiento académico decreciente, o un cambio en la manera en que se otorgan y distribuyen las calificaciones.
- A pesar de los cambios observados en la media, la desviación estándar y los percentiles, los valores mínimos y máximos de las calificaciones se han mantenido constantes en todos los periodos. Esto sugiere que no ha habido un cambio en los límites absolutos de las calificaciones que los estudiantes pueden obtener.
- Aunque el promedio de las notas parciales y finales ha disminuido durante los periodos, sigue siendo relativamente alto, lo que indica que, a pesar de la disminución observada, el rendimiento académico general todavía es considerablemente bueno.

Ahora, una vez finalizada el proceso de análisis mediante estadísticas descriptivas de las variables numéricas del reporte de calificaciones por carrera, es necesario realizar una exploración mediante histogramas.

Para comenzar, a continuación, se podrá visualizar los histogramas de las notas parciales el reporte de calificaciones por carrera, sin embargo, es necesario recordar que esta exploración es realizada de manera informativa y no se tomará en cuenta en el análisis posterior a la presente exploración de variables numéricas.

Ilustración 126

Código para la generación de los histogramas de las notas parciales del reporte de Calificaciones por Carrera

```
# Se crea una figura y ejes
fig, axes = plt.subplots(2, 2, figsize=(12, 12))

periodos = ["2021-01", "2021-02", "2022-01", "2022-02"]

# Se itera sobre los índices de las subfiguras y los datos correspondientes
for i, ax in enumerate(axes.flatten()):
    # Se crea una tabla de frecuencia para el dataset actual
    tabla_frecuencia = pd.cut(numeric_calificaciones[i]['nota_parcial'], bins=np.arange(0, 55, 5)).value_counts(sort=False)

    # Se crea un histograma a partir de la tabla de frecuencia y se dibuja en la subfigura actual
    ax.bar(x=tabla_frecuencia.index.astype(str), height=tabla_frecuencia.values)

    # Se agregan etiquetas y título a la subfigura actual
    ax.set_xlabel("Rangos de notas parciales")
    ax.set_ylabel("Frecuencia")
    ax.set_title(f"Periodo {periodos[i]}")

    # Se giran verticalmente los xticks
    ax.tick_params(axis='x', rotation=90)

# Se ajusta el espacio entre las subfiguras
plt.tight_layout()

# Se muestra el gráfico completo que contiene las subfiguras
plt.show()
```

Ilustración 127

Histogramas del primer y segundo periodo de la variable nota_parcial del reporte de Calificaciones por Carrera

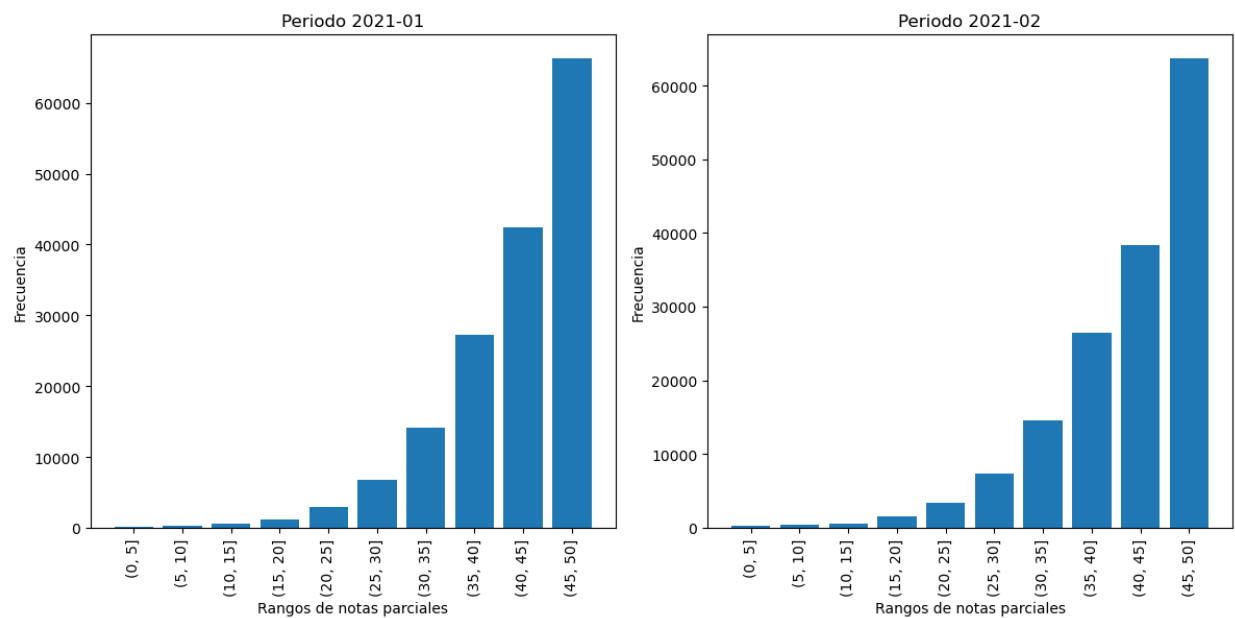
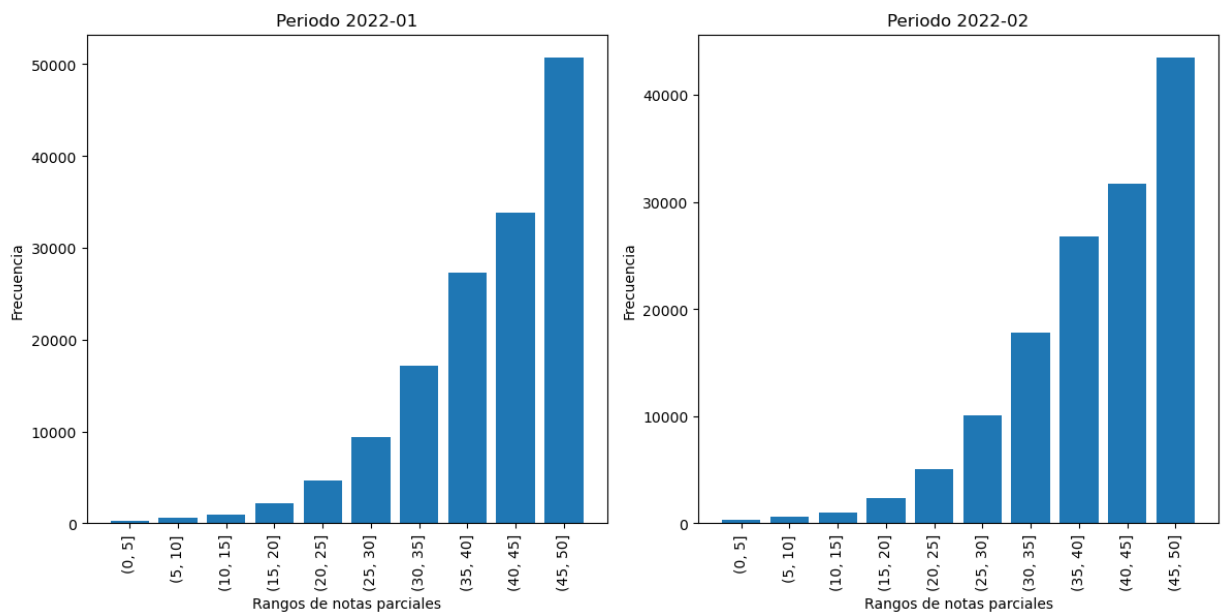


Ilustración 128

Histogramas del tercer y cuarto periodo de la variable nota_parcial del reporte de Calificaciones por Carrera



A partir de estos histogramas se pudo concluir lo siguiente:

- Se puede observar que en todos los periodos se presenta una distribución en forma de "J" o asimétrica, lo que implica que la mayoría de los estudiantes obtienen calificaciones altas o satisfactorias en sus parciales, mientras que hay un grupo reducido de estudiantes con calificaciones muy bajas.

Seguido de la conclusión mencionada, es momento de presentar los histogramas de las calificaciones finales:

Ilustración 129

Código que genera los histogramas de las notas finales del reporte de Calificaciones por Carrera

```
# Se crea una figura y ejes
fig, axes = plt.subplots(2, 2, figsize=(12, 12))

periodos = ["2021-01", "2021-02", "2022-01", "2022-02"]

# Se itera sobre los índices de las subfiguras y los datos correspondientes
for i, ax in enumerate(axes.flatten()):
    # Se crea una tabla de frecuencia para el dataset actual
    tabla_frecuencia = pd.cut(numeric_calificaciones[i]['nota_final'], bins=np.arange(0, 55, 5)).value_counts(sort=False)

    # Se crea un histograma a partir de la tabla de frecuencia y se dibuja en la subfigura actual
    ax.bar(x=tabla_frecuencia.index.astype(str), height=tabla_frecuencia.values)

    # Se agregan etiquetas y título a la subfigura actual
    ax.set_xlabel("Rangos de notas finales")
    ax.set_ylabel("Frecuencia")
    ax.set_title(f"Periodo {periodos[i]}")

    # Se giran verticalmente los xticks
    ax.tick_params(axis='x', rotation=90)

# Se ajusta el espacio entre las subfiguras
plt.tight_layout()

# Se muestra el gráfico completo que contiene las subfiguras
plt.show()
```

Ilustración 130

Histogramas del primer y segundo periodo de la variable nota_final del reporte de Calificaciones por Carrera

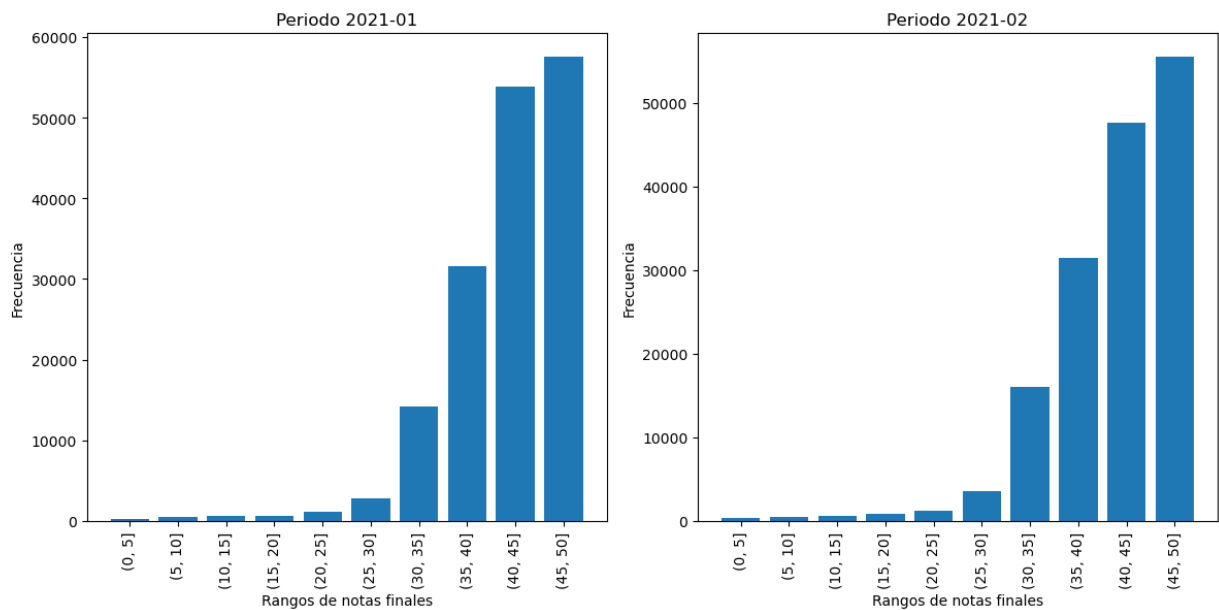
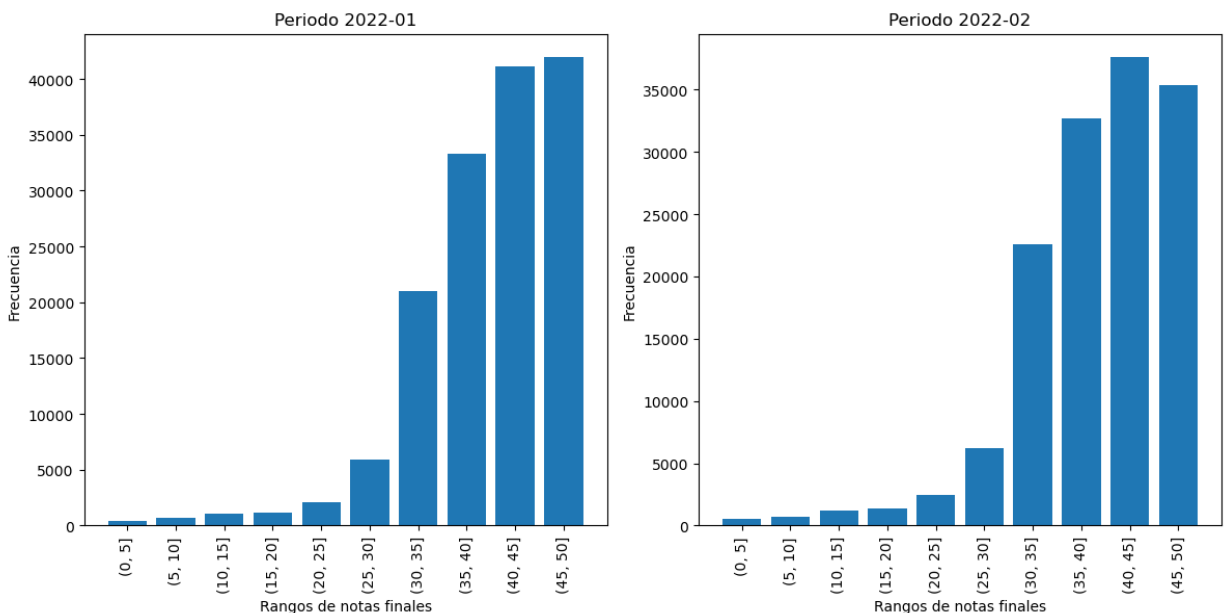


Ilustración 131

Histogramas del tercer y cuarto periodo de la variable nota_final del reporte de Calificaciones por Carrera



A partir de los histogramas visibles, se logró concluir lo siguiente:

- Se observa una diferencia en la distribución de las calificaciones finales en comparación con las calificaciones parciales, concretamente, existe una mayor tendencia de los datos hacia la derecha, lo que indica un gran porcentaje de estudiantes con buenas notas y un reducido grupo con bajas notas. Lo que es evidente al considerar las dimensiones de los datos de los estudiantes en el reporte de calificaciones por carrera y la cantidad que se encuentran en el reporte de asignaturas perdidas.

Consecuente a la especificación de las conclusiones de los histogramas de ambas variables, se puede determinar que una distribución en forma de "J" como las que se presentan pueden ser un indicador de un rendimiento académico desigual dentro de la propia universidad, donde la mayoría obtiene buenos resultados, pero algunos estudiantes están experimentando dificultades significativas. Además, se observa que con cada periodo que pasa, la distribución de estudiantes con buen desempeño va disminuyendo.

Esta observación resalta la importancia de analizar y contar con un modelo predictivo capaz de comprender cuándo un estudiante, considerando otras variables relevantes, tendrá un bajo rendimiento académico. Un modelo de este tipo podría ser de gran utilidad para identificar a tiempo a aquellos estudiantes que necesitan apoyo adicional y ofrecerles las herramientas y recursos necesarios para mejorar su rendimiento. Esto no solo beneficiaría al estudiante individualmente, sino que también contribuiría a elevar el promedio de calificaciones de la universidad en general.

La implementación de un modelo predictivo permitiría una intervención temprana y personalizada, brindando a los estudiantes oportunidades para mejorar su desempeño y maximizar su potencial académico; ya que, al abordar las dificultades de manera proactiva, la

universidad podría trabajar en la reducción de la brecha entre los estudiantes con calificaciones altas y aquellos con calificaciones bajas, promoviendo un mayor equilibrio y éxito académico para todos los estudiantes.

Por otra parte, es momento de analizar las variables numéricas con otro enfoque utilizando gráficos de densidad, para empezar, se desplegará el gráfico de densidad de las notas parciales:

Ilustración 132

Código que genera el gráfico de densidad de las notas parciales del reporte de Calificaciones por Carrera

```
# Generacion de un dataframe con datos por periodo academico
datos_temp = pd.DataFrame()
datos_temp["PERIODO 2021-01"] = numeric_calificaciones[0]['nota_parcial']
datos_temp["PERIODO 2021-02"] = numeric_calificaciones[1]['nota_parcial']
datos_temp["PERIODO 2022-01"] = numeric_calificaciones[2]['nota_parcial']
datos_temp["PERIODO 2022-02"] = numeric_calificaciones[3]['nota_parcial']

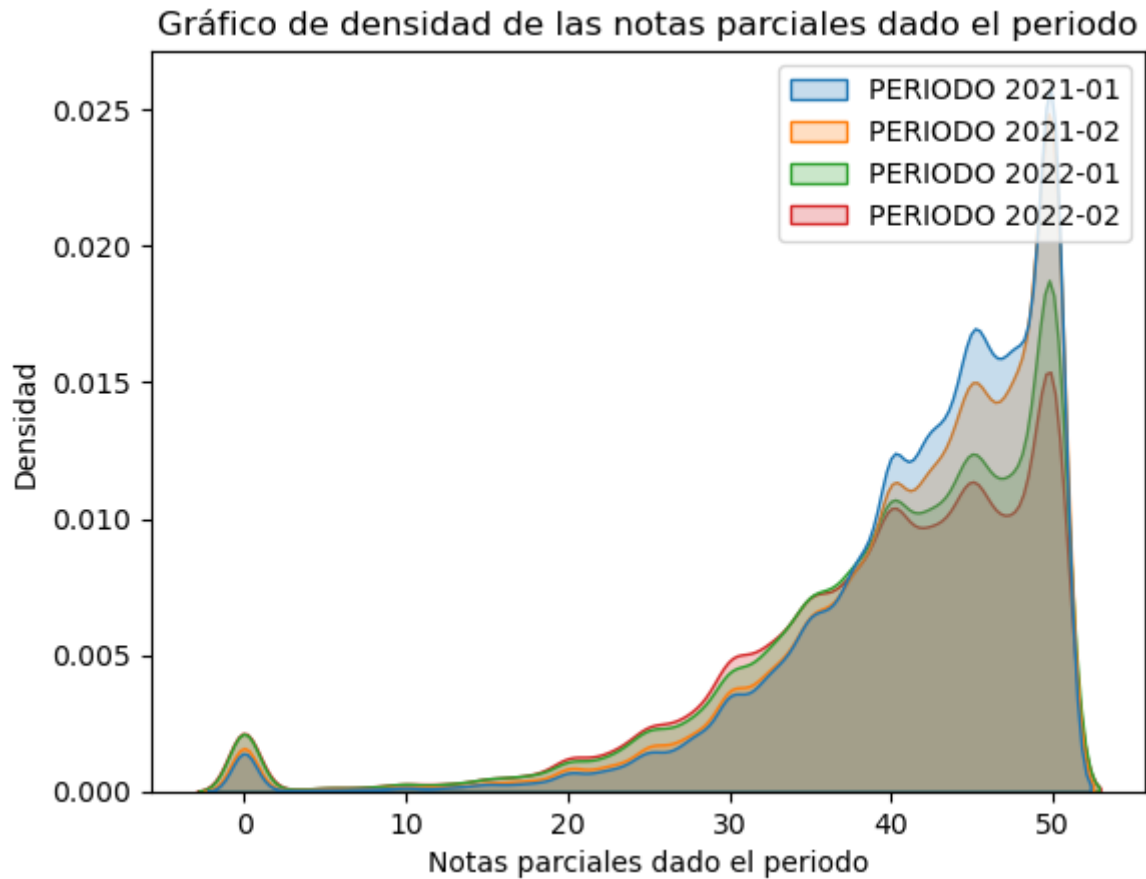
# Crear un gráfico de densidad utilizando Seaborn
sns.kdeplot(data=datos_temp, shade=True)

# Agregar etiquetas y título
plt.xlabel('Notas parciales dado el periodo')
plt.ylabel('Densidad')
plt.title('Gráfico de densidad de las notas parciales dado el periodo')

# Mostrar el gráfico
plt.show()
```

Ilustración 133

Gráfico de densidad de la variable nota_parcial del reporte de Calificaciones por Carrera



A partir del gráfico presentado se puede concluir lo siguiente:

- Se observa que, si bien hubo una disminución significativa en las calificaciones positivas al principio, esta tendencia se recuperó en el último periodo, aunque sigue estando por debajo de los niveles iniciales.
- Aunque la cantidad de distribución varía en cada periodo, se observa que la forma de la distribución es similar. Esto sugiere la posibilidad de que la distribución de calificaciones parciales tenga un patrón inherente.

Seguido de la finalización de las conclusiones del gráfico de densidad de las notas finales, se exhibirá el grafico de densidad de las notas finales del reporte:

Ilustración 134

Código que genera el gráfico de densidad de las notas finales del reporte de Calificaciones por Carrera

```
# Generación de varios dataframes para generar Los gráficos (y permitir mover los identificadores de los gráficos)
datos_temp1 = numeric_calificaciones[0]['nota_final']
datos_temp2 = numeric_calificaciones[1]['nota_final']
datos_temp3 = numeric_calificaciones[2]['nota_final']
datos_temp4 = numeric_calificaciones[3]['nota_final']

# Crear un gráfico de densidad utilizando Seaborn
sns.kdeplot(data=datos_temp1, shade=True, label="PERIODO 2021-01")
sns.kdeplot(data=datos_temp2, shade=True, label="PERIODO 2021-02")
sns.kdeplot(data=datos_temp3, shade=True, label="PERIODO 2022-01")
sns.kdeplot(data=datos_temp4, shade=True, label="PERIODO 2022-02")

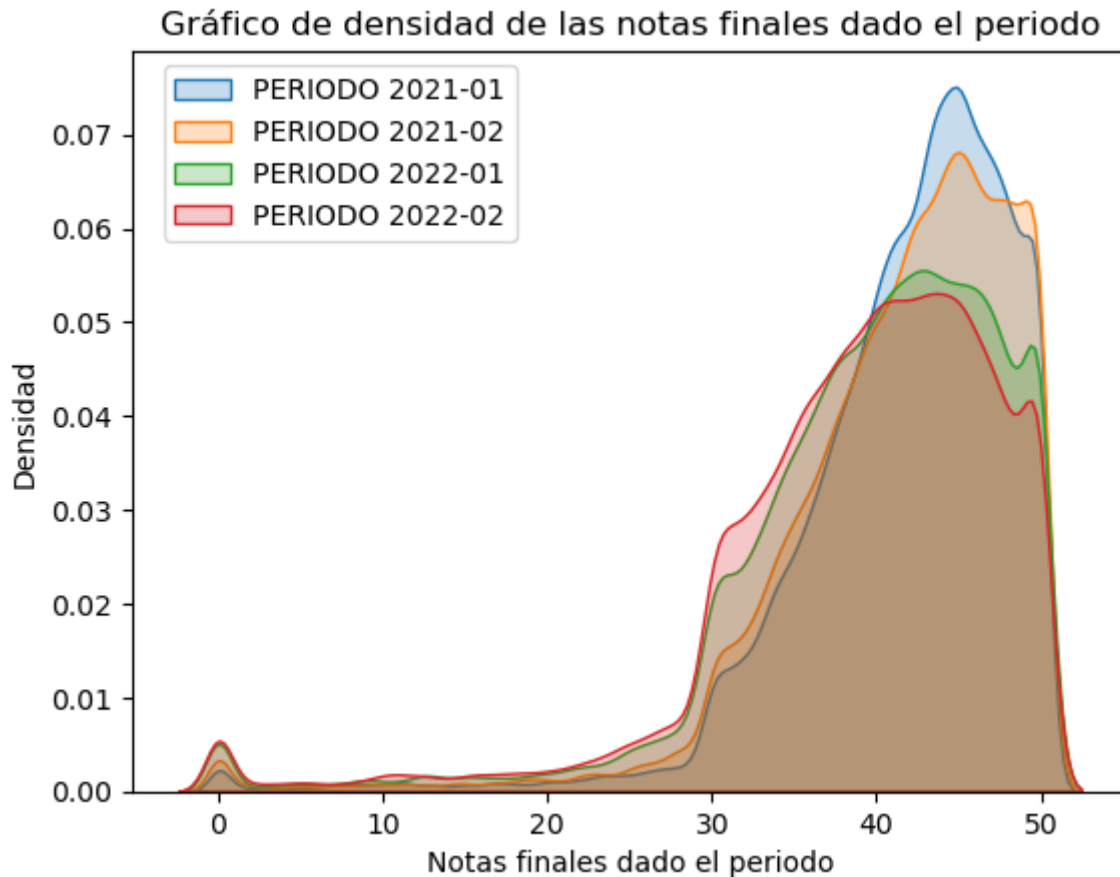
# Agregar etiquetas y título
plt.xlabel('Notas finales dado el periodo')
plt.ylabel('Densidad')
plt.title('Gráfico de densidad de las notas finales dado el periodo')

# Se mueve la ubicación de las descripciones de los gráficos
plt.legend(bbox_to_anchor=(.40, 1))

# Mostrar el gráfico
plt.show()
```

Ilustración 135

Gráfico de densidad de la variable nota_final del reporte de Calificaciones por Carrera



A partir de este gráfico, es posible destacar las siguientes conclusiones:

- Es evidente que a medida que avanzan los periodos, la distribución de calificaciones muestra una reducción progresiva en las calificaciones altas y un aumento correspondiente en las calificaciones bajas.
- A pesar de este cambio en la distribución, se observa que las gráficas siguen un patrón de crecimiento similar en todos los periodos.

Finalmente, el último análisis que se realizará sobre las variables numéricas del reporte de calificaciones por carrera son los diagramas de caja y bigotes, en esta ocasión se realizará

sobre ambas variables dado que ambas son continuas. Para empezar, se revelará el diagrama de las notas parciales:

Ilustración 136

Código que genera los diagramas de caja y bigotes de las notas parciales del reporte de Calificaciones por Carrera

```
plt.figure(figsize=(10, 8))

plt.subplot(221)
sns.boxplot(y='nota_parcial', data=numeric_calificaciones[0])
plt.title("Diagrama de calificaciones parciales - 2021-01")
plt.ylabel('Notas parciales')

plt.subplot(222)
sns.boxplot(y='nota_parcial', data=numeric_calificaciones[1])
plt.title("Diagrama de calificaciones parciales - 2021-02")
plt.ylabel('Notas parciales')

plt.subplot(223)
sns.boxplot(y='nota_parcial', data=numeric_calificaciones[2])
plt.title("Diagrama de calificaciones parciales - 2022-01")
plt.ylabel('Notas parciales')

plt.subplot(224)
sns.boxplot(y='nota_parcial', data=numeric_calificaciones[3])
plt.title("Diagrama de calificaciones parciales - 2022-02")
plt.ylabel('Notas parciales')

plt.tight_layout()
plt.show()
```

Ilustración 137

Diagrama de caja y bigotes del primer y segundo periodo de la variable nota_parcial del reporte de Calificaciones por Carrera

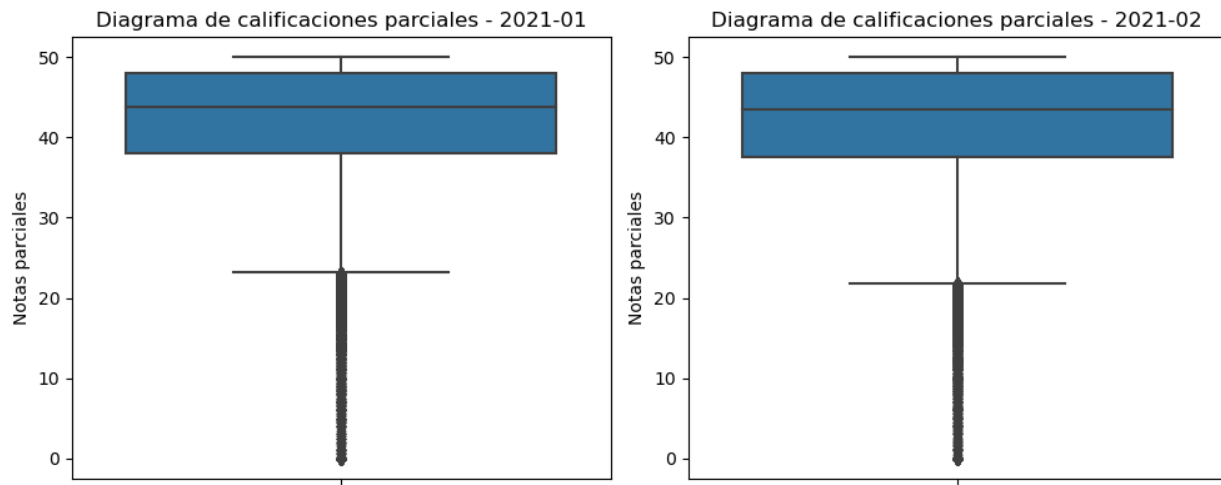
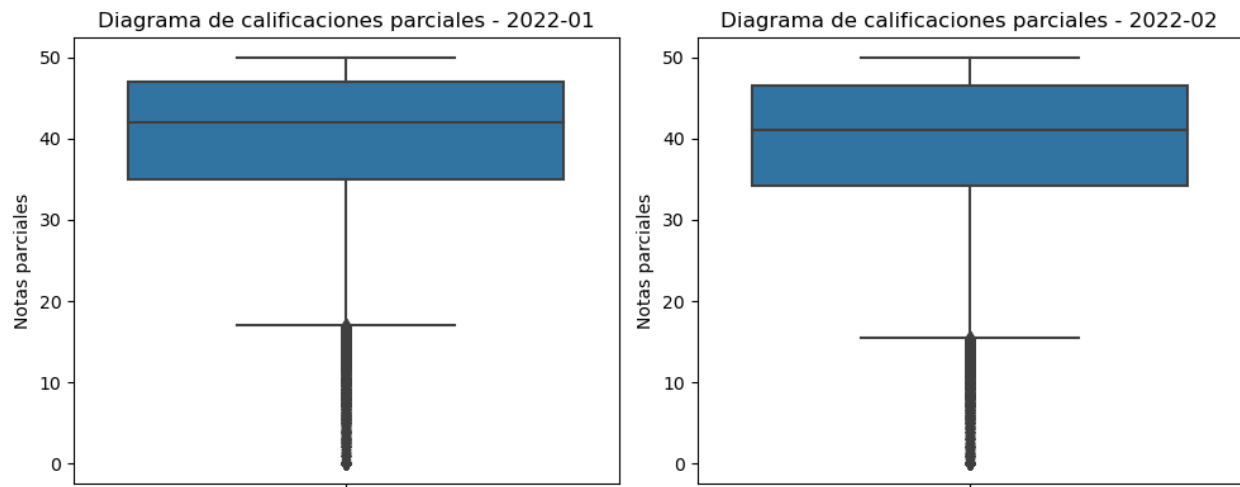


Ilustración 138

Diagrama de caja y bigotes del tercer y cuarto periodo de la variable nota_parcial del reporte de Calificaciones por Carrera



A partir de los diagramas presentados, se puede concluir lo siguiente:

- Se puede observar la presencia de varios valores que podrían considerarse atípicos, pero al encontrarse dentro del rango posible para calificaciones reales, no se consideran propiamente como valores atípicos.

- Además, se nota un comportamiento bastante similar en la distribución de calificaciones entre los diferentes periodos.

En segundo y último lugar, se llevará a cabo el análisis por diagrama de caja y bigotes de las calificaciones finales a continuación:

Ilustración 139

Código que genera los diagramas de caja y bigotes de las calificaciones finales del reporte de Calificaciones por Carrera

```
plt.figure(figsize=(10, 8))

plt.subplot(221)
sns.boxplot(y='nota_final', data=numeric_calificaciones[0])
plt.title("Diagrama de calificaciones finales - 2021-01")
plt.ylabel('Notas parciales')

plt.subplot(222)
sns.boxplot(y='nota_final', data=numeric_calificaciones[1])
plt.title("Diagrama de calificaciones finales - 2021-02")
plt.ylabel('Notas finales')

plt.subplot(223)
sns.boxplot(y='nota_final', data=numeric_calificaciones[2])
plt.title("Diagrama de calificaciones finales - 2022-01")
plt.ylabel('Notas finales')

plt.subplot(224)
sns.boxplot(y='nota_final', data=numeric_calificaciones[3])
plt.title("Diagrama de calificaciones finales - 2022-02")
plt.ylabel('Notas finales')

plt.tight_layout()
plt.show()
```


Ilustración 140

Diagramas de caja y bigotes del primer y segundo periodo de la variable nota_final del reporte de Calificaciones por Carrera

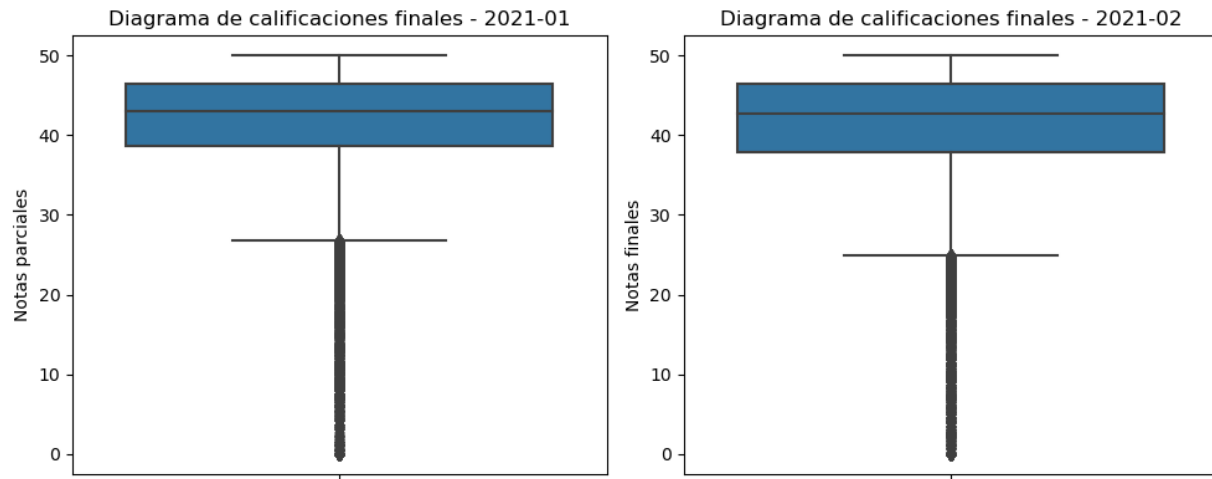
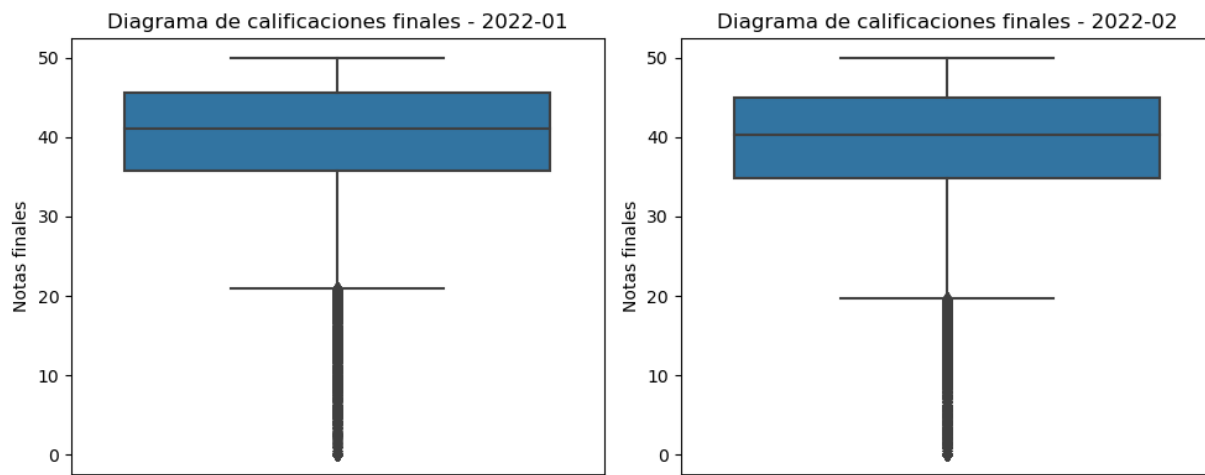


Ilustración 141

Diagramas de caja y bigotes del tercer y cuarto periodo de la variable nota_final del reporte de Calificaciones por Carrera



A partir de los diagramas realizados se puede concluir lo siguiente:

- Se observa la presencia de posibles valores atípicos, sin embargo, al pertenecer al rango de valores posibles para calificaciones, no se consideran propiamente como atípicos.

- Se evidencia una marcada similitud entre las distribuciones de calificaciones en todos los periodos.

Finalmente se ha finalizado la exploración de variables numéricas del reporte de calificaciones por carrera, a partir de todo lo realizado, por lo que es necesario mencionar que se logró la identificación de las siguientes variables de interés de este reporte con lo que respecta a variables numéricas:

- nota_final

Esto quiere decir que la variable de interés debe seguir siendo considerada para los análisis consiguientes, sin embargo, esto no quiere decir que la variable nota_parcial no sea de interés general para el rendimiento académico, únicamente no es relevante para el enfoque de esta investigación.

En última instancia, se realizará el análisis de las variables numéricas del reporte de becas comenzando por las estadísticas descriptivas correspondientes a su primer periodo:

Ilustración 142

Estadísticas Descriptivas del primer periodo del reporte de Becas

```
numeric_becas[0].describe()
```

	NUM_MATERIAS_SEGUNDA_MATRICULA	NUM_MATERIAS_TERCERA_MATRICULA
count	9399.000000	9399.000000
mean	0.108522	0.013938
std	0.450025	0.148474
min	0.000000	0.000000
25%	0.000000	0.000000
50%	0.000000	0.000000
75%	0.000000	0.000000
max	6.000000	4.000000

A partir de las presentes estadísticas descriptivas del primer periodo del reporte de becas, se puede observar lo siguiente:

- Se puede observar que posiblemente no existan valores nulos.
- Se puede observar que al existir un promedio de 0.11 materias en segunda matrícula y un promedio de 0.01 de materias en tercera matrícula, es evidente que aquellas personas que reprueban una materia tienen una menor probabilidad de reprobarla una vez más.
- En ambos casos, al observar que el 75% de los registros tienen un valor de 0 quiere decir que no es un porcentaje alto el que presenta reprobaciones en asignaturas.
- Por último, no se observan valores atípicos, ya que los valores mínimos son 0 y los máximos son 6 y 4, lo cual está dentro del rango real posible.

A continuación, se podrá visualizar las estadísticas descriptivas del reporte correspondiéndose a su tercer periodo:

Ilustración 143

Estadísticas Descriptivas del segundo periodo del reporte de Becas

```
numeric_becas[1].describe()
```

	NUM_MATERIAS_SEGUNDA_MATRICULA	NUM_MATERIAS_TERCERA_MATRICULA
count	8564.000000	8564.000000
mean	0.097618	0.014012
std	0.406991	0.145113
min	0.000000	0.000000
25%	0.000000	0.000000
50%	0.000000	0.000000
75%	0.000000	0.000000
max	6.000000	3.000000

A partir de estas estadísticas se pueden observar los siguientes aspectos:

- Se puede observar que posiblemente no existan valores nulos.
- Se puede observar que al existir un promedio de 0.09 materias en segunda matrícula y un promedio de 0.01 de materias en tercera matrícula, es evidente que aquellas personas que reprueban una materia tienen una menor probabilidad de reprobarla una vez más.
- En ambos casos, al observar que el 75% de los registros tienen un valor de 0 quiere decir que no es un porcentaje alto el que presenta reprobaciones en asignaturas.

- Por último, no se observan valores atípicos, ya que los valores mínimos son 0 y los máximos son 6 y 3, lo cual está dentro del rango real posible.

En tercer lugar, se exhibirán las estadísticas descriptivas correspondientes al tercer periodo del reporte de becas:

Ilustración 144

Estadísticas Descriptivas del tercer periodo del reporte de Becas

```
numeric_becas[2].describe()
```

	NUM_MATERIAS_SEGUNDA_MATRICULA	NUM_MATERIAS_TERCERA_MATRICULA
count	8336.000000	8336.000000
mean	0.142514	0.019914
std	0.520105	0.182200
min	-1.000000	-1.000000
25%	0.000000	0.000000
50%	0.000000	0.000000
75%	0.000000	0.000000
max	6.000000	4.000000

A partir de las presentes estadísticas descriptivas, se pueden observar los siguientes comportamientos:

- Se puede observar que posiblemente no existan valores nulos.
- Se puede observar que al existir un promedio de 0.15 materias en segunda matrícula y un promedio de 0.01 de materias en tercera matrícula, es evidente que aquellas personas que reprueban una materia tienen una menor probabilidad de reprobarla una vez más.

- En ambos casos, al observar que el 75% de los registros tienen un valor de 0 quiere decir que no es un porcentaje alto el que presenta reprobaciones en asignaturas.
- Por último, no se observan valores atípicos, ya que los valores mínimos son 0 y los máximos son 6 y 4, lo cual está dentro del rango real posible.

Finalmente, es momento de presentar las estadísticas descriptivas del cuarto y último periodo del reporte de becas:

Ilustración 145

Estadísticas Descriptivas del cuarto periodo del reporte de Becas

```
numeric_becas[3].describe()
```

	NUM_MATERIAS_SEGUNDA_MATRICULA	NUM_MATERIAS_TERCERA_MATRICULA
count	8255.000000	8255.000000
mean	0.179528	0.019624
std	0.567894	0.176394
min	0.000000	0.000000
25%	0.000000	0.000000
50%	0.000000	0.000000
75%	0.000000	0.000000
max	6.000000	5.000000

A partir de las estadísticas descriptivas se pueden observar los siguientes comportamientos:

- Se puede observar que posiblemente no existan valores nulos.
- Se puede observar que al existir un promedio de 0.17 materias en segunda matrícula y un promedio de 0.01 de materias en tercera matrícula, es evidente que aquellas

personas que reprueban una materia tienen una menor probabilidad de reprobirla una vez más.

- En ambos casos, al observar que el 75% de los registros tienen un valor de 0 quiere decir que no es un porcentaje alto el que presenta reprobaciones en asignaturas.
- Por último, no se observan valores atípicos, ya que los valores mínimos son 0 y los máximos son 6 y 5, lo cual está dentro del rango real posible.

En resumen, a causa del análisis de todas las estadísticas descriptivas respectivas del reporte de becas, se definieron las siguientes conclusiones:

- No parece haber evidencia de valores nulos en los períodos observados, lo cual indica que los datos están completos y que todos los estudiantes tienen registros de matrículas.
- A lo largo de los períodos observados, se puede notar una consistencia en los promedios de materias en segunda y tercera matrícula. Esto sugiere que la proporción de estudiantes que reprobaron y necesitaron retomar asignaturas se mantiene bastante constante.
- El hecho de que los valores de la tercera matrícula sean consistentemente menores indica que aquellos estudiantes que necesitan retomar una asignatura en la segunda matrícula tienden a tener una menor probabilidad de reprobirla nuevamente. Sin embargo, este es un hallazgo preliminar y se requiere un análisis más detallado para confirmarlo.

- El 75% de los registros muestran un valor de 0 en la segunda y tercera matrícula, lo que indica que la gran mayoría de los estudiantes no reprueba las asignaturas. Esto es un indicativo de un rendimiento académico generalmente sólido entre los estudiantes.
- A pesar de la variabilidad en los promedios, los valores mínimos y máximos de las materias en segunda y tercera matrícula se mantienen dentro de los límites esperados. Esto sugiere que no hay valores atípicos que podrían sesgar los resultados del análisis.

Una vez concretada la finalización del análisis exploratorio haciendo uso de la técnica de estadísticas descriptivas, es momento de realizar un análisis con un diferente enfoque, en esta ocasión en semejanza con los anteriores reportes evaluados se utilizarán histogramas. A continuación, se comenzará visualizando un histograma de la cantidad de segundas matrículas que posee un determinado estudiante.

Ilustración 146

Código que genera los histogramas de la cantidad de segundas matrículas que poseen los estudiantes del reporte de Becas

```
# Se crea una figura y ejes
fig, axes = plt.subplots(2, 2, figsize=(12, 12))

periodos = ["2021-01", "2021-02", "2022-01", "2022-02"]

# Se itera sobre los índices de las subfiguras y los datos correspondientes
for i, ax in enumerate(axes.flatten()):
    # Se crea una tabla de frecuencia para el dataset actual
    tabla_frecuencia = pd.cut(numeric_becas[i]['NUM_MATERIAS_SEGUNDA_MATRICULA'], bins=np.arange(0, 10, 1)).value_counts(sort=False)

    # Se crea un histograma a partir de la tabla de frecuencia y se dibuja en la subfigura actual
    ax.bar(x=tabla_frecuencia.index.astype(str), height=tabla_frecuencia.values)

    # Se agregan etiquetas y título a la subfigura actual
    ax.set_xlabel("Rangos de cantidad de segundas matrículas")
    ax.set_ylabel("Frecuencia")
    ax.set_title(f"Periodo {periodos[i]}")

    # Se giran verticalmente los xticks
    ax.tick_params(axis='x', rotation=90)

# Se ajusta el espacio entre las subfiguras
plt.tight_layout()

# Se muestra el gráfico completo que contiene las subfiguras
plt.show()
```


Ilustración 147

Histogramas del primer y segundo periodo de la variable
NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas

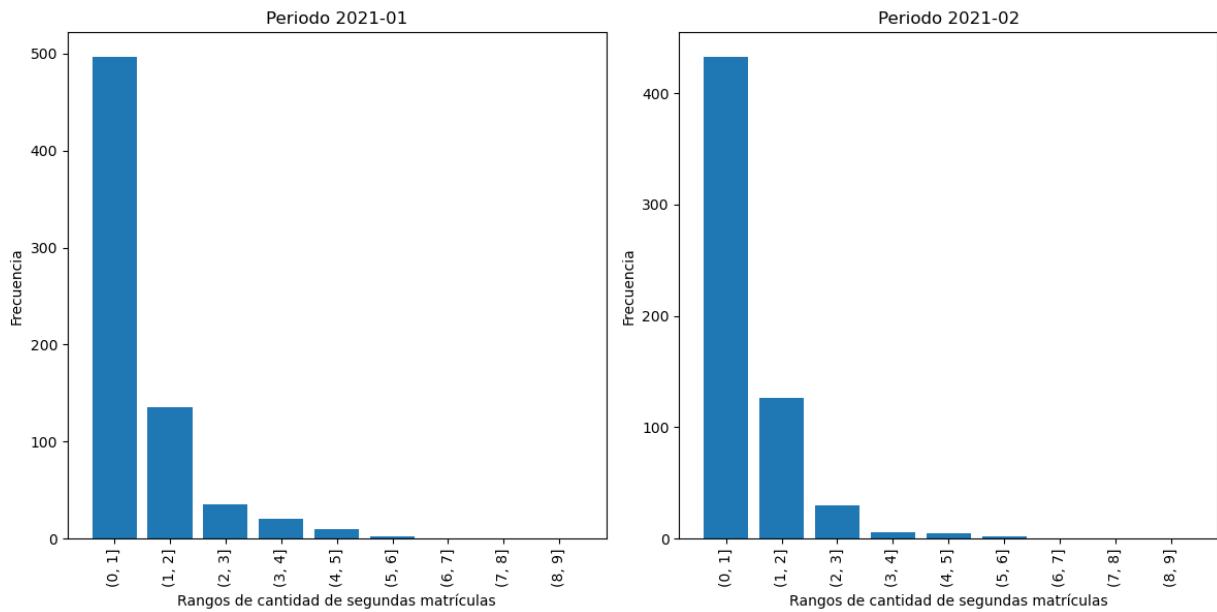
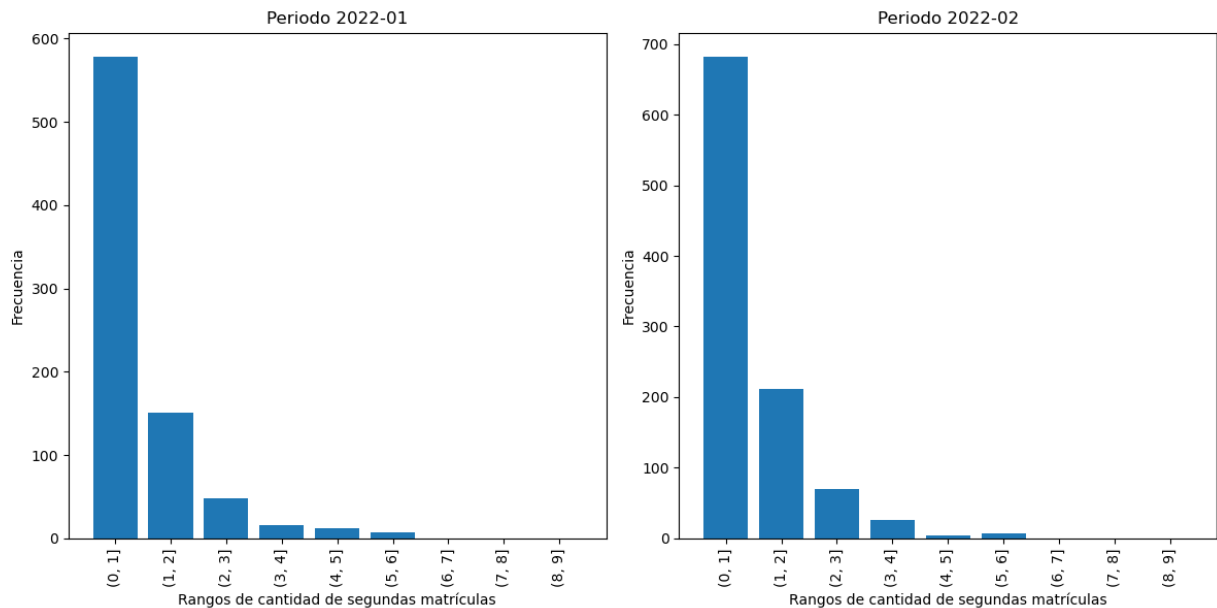


Ilustración 148

Histogramas del tercer y cuarto periodo de la variable
NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas



A partir de los histogramas realizados de la cantidad de segundas matrículas que posee un estudiante, se ha concluido que se puede observar claramente un patrón en el que, si un estudiante reprueba una asignatura, es menos probable que repruebe otra asignatura adicional; lo que indica que existe una tendencia a que las calificaciones sean consistentes entre las asignaturas, y que el rendimiento en una materia puede ser un indicador del rendimiento en otras materias.

Sin embargo, es importante destacar que este patrón se elimina al momento de sobrepasar 4 materias reprobadas. En otras palabras, si un estudiante reprueba 5 materias, no se puede predecir con exactitud si tiene una mayor o menor probabilidad de reprobado una sexta materia. Esto sugiere que existe una variabilidad significativa en el rendimiento académico de los estudiantes y que hay casos en los que un bajo rendimiento en múltiples asignaturas puede no seguir la misma tendencia de las calificaciones individuales.

Por otra parte, es el tiempo adecuado de presentar los histogramas de la cantidad de terceras matrículas que poseen los estudiantes del reporte de becas:

Ilustración 149

Código que genera los histogramas de la cantidad de terceras matrículas que poseen los estudiantes del reporte de Becas

```
# Se crea una figura y ejes
fig, axes = plt.subplots(2, 2, figsize=(12, 12))

periodos = ["2021-01", "2021-02", "2022-01", "2022-02"]

# Se itera sobre los índices de las subfiguras y los datos correspondientes
for i, ax in enumerate(axes.flatten()):
    # Se crea una tabla de frecuencia para el dataset actual
    tabla_frecuencia = pd.cut(numeric_becas[i]['NUM_MATERIAS_TERCERA_MATRICULA'], bins=np.arange(0, 10, 1)).value_counts(sort=False)

    # Se crea un histograma a partir de la tabla de frecuencia y se dibuja en la subfigura actual
    ax.bar(x=tabla_frecuencia.index.astype(str), height=tabla_frecuencia.values)

    # Se agregan etiquetas y título a la subfigura actual
    ax.set_xlabel("Rangos de cantidad de terceras matrículas")
    ax.set_ylabel("Frecuencia")
    ax.set_title(f"Período {periodos[i]}")

    # Se giran verticalmente los xticks
    ax.tick_params(axis='x', rotation=90)

# Se ajusta el espacio entre las subfiguras
plt.tight_layout()

# Se muestra el gráfico completo que contiene las subfiguras
plt.show()
```

Ilustración 150

Histogramas del primer y segundo periodo de la variable NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas

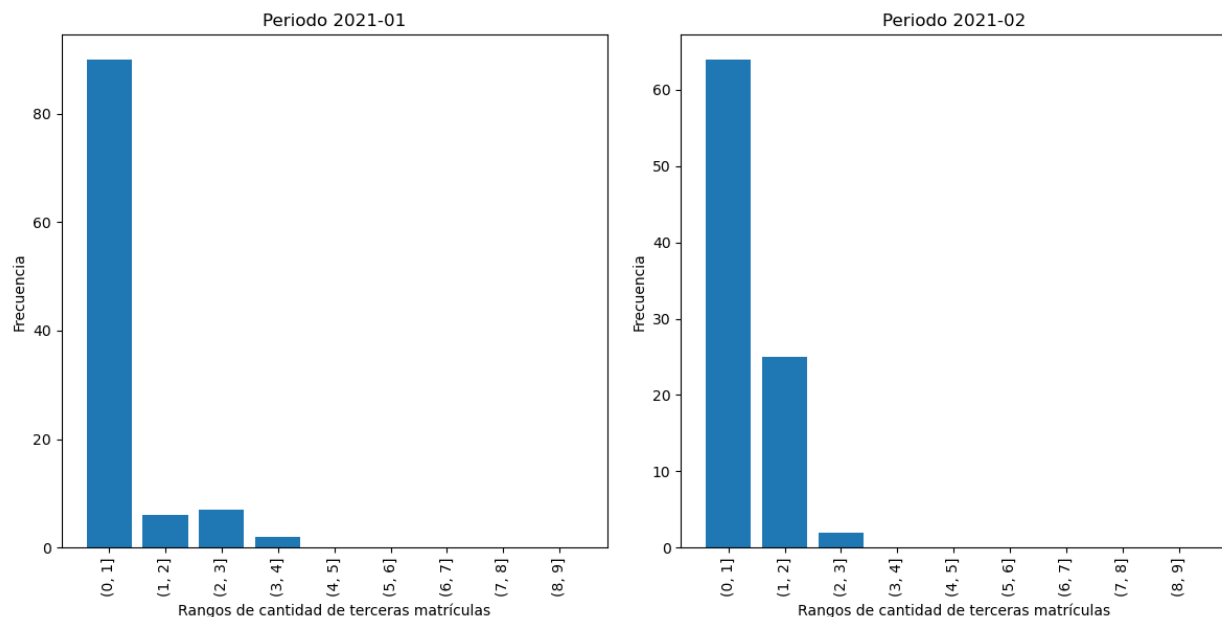
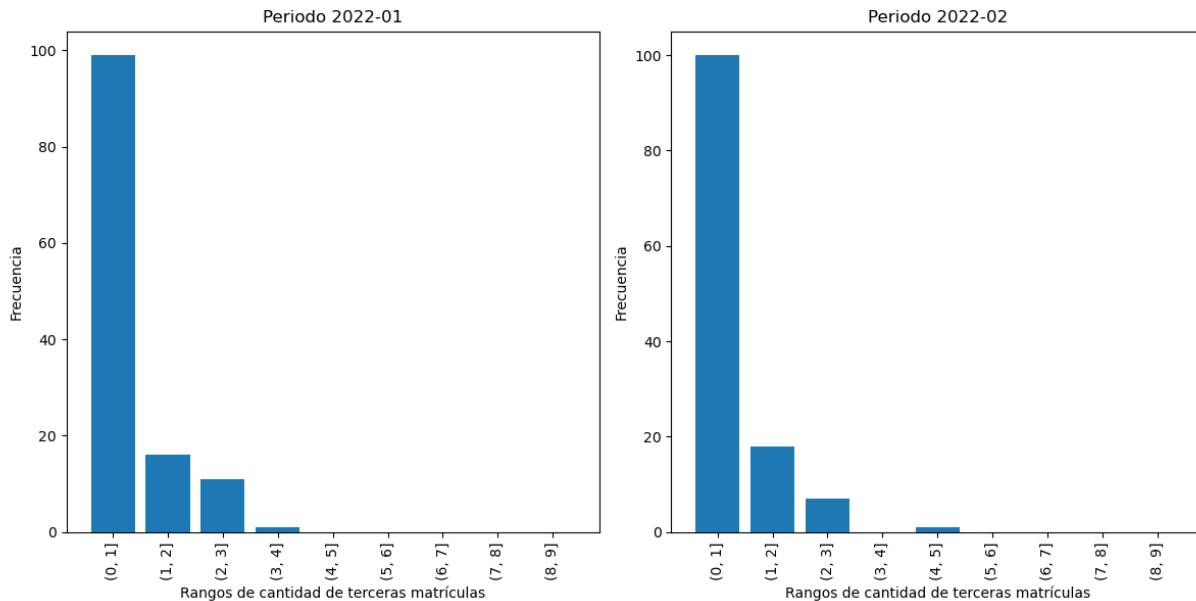


Ilustración 151

Histogramas del tercer y cuarto periodo de la variable
NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas



A partir de los histogramas exhibidos, se puede concluir que al igual que la variable anterior, se destaca que es menos probable que un estudiante que obtuvo una tercera matrícula en una asignatura tenga otra tercera matrícula en otra asignatura. Sin embargo, no se evidencia un patrón claro de comportamiento a partir del momento en que un estudiante tiene dos o más asignaturas con tercera matrícula.

Por lo que, se sugiere que existe cierta consistencia en el rendimiento académico de un estudiante en lo que respecta a las terceras matrículas. Es probable que aquellos que han obtenido una tercera matrícula en una asignatura se esfuercen por mejorar su desempeño en las siguientes asignaturas y evitar repetir el mismo resultado. Sin embargo, a medida que aumenta el número de asignaturas con tercera matrícula, no se puede determinar con certeza si este patrón se mantiene o si el estudiante logrará mejorar su rendimiento.

En diferente instancia, es momento de enfocar el análisis al uso de gráficos de densidad con la finalidad de conseguir una visión más precisa de las distribuciones a través del tiempo, es decir, a través de los diferentes periodos del reporte de becas. En un inicio se revelará el gráfico de densidad de la cantidad de asignaturas con segunda matrícula que poseen los estudiantes:

Ilustración 152

Código que genera el gráfico de densidad de la cantidad de segundas matrículas que poseen los estudiantes en el reporte de Becas

```
# Generacion de un dataframe con datos por periodo academico
datos_temp = pd.DataFrame()
datos_temp["PERIODO 2021-01"] = numeric_becas[0]['NUM_MATERIAS_SEGUNDA_MATRICULA']
datos_temp["PERIODO 2021-02"] = numeric_becas[1]['NUM_MATERIAS_SEGUNDA_MATRICULA']
datos_temp["PERIODO 2022-01"] = numeric_becas[2]['NUM_MATERIAS_SEGUNDA_MATRICULA']
datos_temp["PERIODO 2022-02"] = numeric_becas[3]['NUM_MATERIAS_SEGUNDA_MATRICULA']

# Crear un gráfico de densidad utilizando Seaborn
sns.kdeplot(data=datos_temp, multiple="stack")

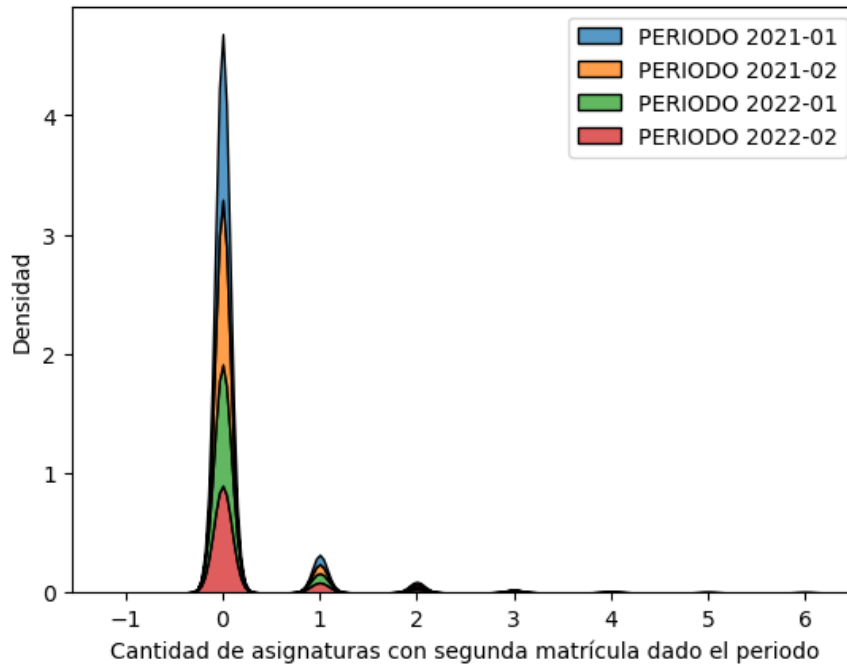
# Agregar etiquetas y título
plt.xlabel('Cantidad de asignaturas con segunda matrícula dado el periodo')
plt.ylabel('Densidad')
plt.title('Gráfico de densidad de la cantidad de asignaturas con segunda matrícula dado el periodo')

# Mostrar el gráfico
plt.show()
```

Ilustración 153

Gráfico de densidad de la variable NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas

Gráfico de densidad de la cantidad de asignaturas con segunda matrícula dado el periodo



A partir de los gráficos visualizados primero hay que realizar una distinción con respecto al código, es claro que es diferente al empleado en anteriores generaciones de gráficos de densidad y su resultado es de igual forma evidente la manera en que se diferencia; su razón es debido a que, si se utilizase la misma forma de visualización que los anteriores gráficos de densidad, sería más intrincado su análisis debido a lo solapado que se encuentran sus datos.

Sin embargo, a pesar de la diferencia descrita, se pudo concluir lo siguiente:

- Se puede observar que a medida que los periodos avanzan, se registra una menor cantidad de estudiantes con 0 segundas matrículas. Esto puede deberse a la presencia de valores nulos que no reflejan asignaturas reprobadas. No obstante, es necesario realizar un análisis más exhaustivo de los datos.

- Se puede observar que a medida que los periodos avanzan, se evidencia una disminución en la cantidad de estudiantes con un menor número de segundas matrículas. Esta tendencia podría ser atribuida a la pérdida de becas por parte de los estudiantes o a la reprobación de otras asignaturas.
- El mismo patrón mencionado en el punto anterior se observa también en las siguientes cantidades de segundas matrículas.

A causa de la finalización del análisis por gráfico de densidad de la cantidad de asignaturas con segundas matrículas que poseen los estudiantes del reporte, es momento idóneo para la evaluación de la cantidad de terceras matrículas que poseen los estudiantes:

Ilustración 154

Código para generar los gráficos de calor de la cantidad de terceras matrículas que poseen los estudiantes del reporte de Becas

```
# Generacion de un dataframe con datos por periodo academico
datos_temp = pd.DataFrame()
datos_temp["PERIODO 2021-01"] = numeric_becas[0]['NUM_MATERIAS_TERCERA_MATRICULA']
datos_temp["PERIODO 2021-02"] = numeric_becas[1]['NUM_MATERIAS_TERCERA_MATRICULA']
datos_temp["PERIODO 2022-01"] = numeric_becas[2]['NUM_MATERIAS_TERCERA_MATRICULA']
datos_temp["PERIODO 2022-02"] = numeric_becas[3]['NUM_MATERIAS_TERCERA_MATRICULA']

# Crear un gráfico de densidad utilizando Seaborn
sns.kdeplot(data=datos_temp, multiple="stack")

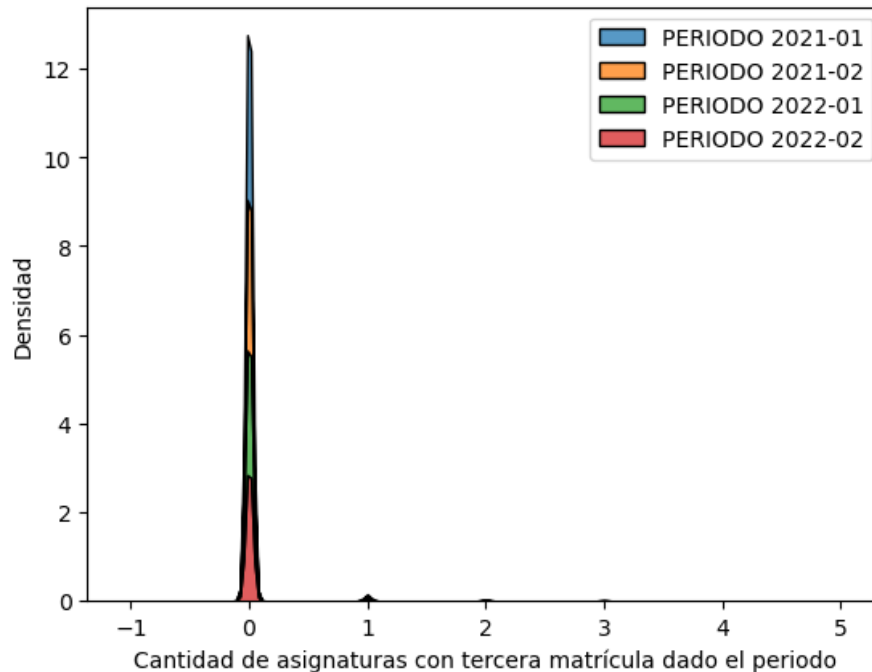
# Agregar etiquetas y título
plt.xlabel('Cantidad de asignaturas con tercera matrícula dado el periodo')
plt.ylabel('Densidad')
plt.title('Gráfico de densidad de la cantidad de asignaturas con tercera matrícula dado el periodo')

# Mostrar el gráfico
plt.show()
```

Ilustración 155

Gráfico de densidad de la variable NUM_MATERIAS_TERCERA_MATIRCULA del reporte de Becas

Gráfico de densidad de la cantidad de asignaturas con tercera matrícula dado el periodo



A partir del gráfico de densidad presentado se pueden destacar observaciones similares a la gráfica de densidad de las segundas matrículas, es decir, se presentan observaciones similares.

En síntesis final de los análisis exploratorios de las variables numéricas del reporte de becas y de todos los demás reportes, lo lógico debería ser continuar con el uso de diagramas de caja y bigotes para analizar las variables numéricas del último reporte, sin embargo, debido a que todas sus variables son discretas, no es prudente la realización de dichos diagramas, por lo que, de manera concluyente, se puede mencionar que las variables de interés del reporte de becas son las dos variables numéricas existentes del mismo, es decir, NUM_MATERIAS_SEGUNDA_MATRICULA y NUM_MATERIAS_TERCERA_MATRICULA.

Como resultado final del largo y arduo proceso de exploración de variables numéricas de los tres diferentes reportes que se han logrado obtener para este trabajo de titulación, se puede cerciorar que se han logrado rescatar diferentes observaciones y conclusiones que reflejan la importancia del resultado de este trabajo de titulación, es decir, de un modelo de predicción del rendimiento académico; sin embargo, el momento apropiado para resaltar dichos resultados será en el capítulo 6 de este documento, por lo que, si el lector se encuentra ávido de conocer a detalle sobre los resultados hallados, es cordialmente invitado a revisarlos en el mencionado capítulo.

4.1.11. Exploración de variables categóricas

En esta sección, una vez finalizado el proceso de exploración de todas las variables numéricas que llevan consigo todos los reportes de datos utilizados en esta investigación, ha llegado al preciso momento de dirigir la atención de la investigación a la exploración de variables categóricas, sin embargo, es prudente contemplar lo qué son las variables categóricas nuevamente y mencionar que técnicas serán aplicadas para su debido análisis en este trabajo de titulación.

Por lo que, las variables categóricas son aquellas que expresan un atributo cualitativo como el color de cabello, color de ojos, religión, película favorita, género, entre otros; además, los valores de este tipo de variables no poseen un orden numérico. (Ziemer, s. f., pág. 28). Es decir, las variables categóricas son aquellas que expresan cualidades de los registros, en esta ocasión, cualidades de los estudiantes de la Pontificia Universidad Católica del Ecuador como su facultad, género, etnia, etc.

A partir de esta premisa, se realizará un análisis exploratorio de las variables categóricas que los reportes poseen, para esto, se hará uso de estadísticas descriptivas, gráficos de barras y de pastel; sin embargo, es necesario comentar que no se utilizarán los gráficos de barras y de pastel

en las mismas variables, es decir, se decidirá uno u otro, ya que, ambos gráficos de pastel y de barras pueden ser métodos efectivos para visualizar datos cualitativos, sin embargo, los gráficos de barras son mejores cuando existen más que unas pocas categorías y para comparar dos o más distribuciones. De todas maneras, es necesario tener presente que hay que evitar crear gráficos engañosos. (Lane, s. f., pág. 74). En otras palabras, con el objetivo de presentar gráficos con información útil, se determinará que método es mejor para visualizar los datos de determinadas variables.

Teniendo todo esto en cuenta, es momento de comenzar a explorar las variables categóricas de todos los reportes utilizados para este trabajo de titulación, sin embargo, es crítico realizar una distinción de que variables serán utilizadas para el análisis y cuáles no. Las variables que no serán sujetas a una exploración más allá del análisis por estadísticas descriptivas, son aquellas relacionadas a la identificación de los estudiantes y aquellas relacionadas a la identificación del periodo, esto se debe a dado que en el enfoque de la investigación, el modelo de machine planteado no tiene como objetivo predecir el rendimiento académico basado en el rendimiento histórico de un estudiante en todas sus asignaturas a lo largo de su vida académica, sino, realizar la predicción basándose en rendimiento histórico de todos los estudiantes en las diversas áreas de asignaturas; por lo que, realizar un análisis sobre esta variable resulta innecesaria, sin embargo, la razón de su permanencia en los marcos de datos y su relevancia para la investigación, es su futura utilidad para la unión de los diferentes marcos de datos correspondientes a cada periodo de cada reporte de datos.

Por otra parte, un análisis realizado para las variables relacionada a la identificación de los periodos resultaría irrelevante, puesto que, todos los datos de cada determinado periodo corresponden únicamente a su periodo. Dicho esto, los análisis de la exploración de variables

categorías comenzarán con el reporte de asignaturas perdidas permitiendo la visualización de sus estadísticas descriptivas por periodo seguidas de sus conclusiones respectivas.

Por lo que, como se mencionó, en primera instancia se visualizará las estadísticas descriptivas del primer periodo del reporte de asignaturas perdidas:

Ilustración 156

Estadísticas descriptivas del primer periodo del reporte de Asignaturas Perdidas

```
categoric_perdidas[0].describe()
```

	IDD	PERIODO	PLAN_E	DESC_FACULTAD	DESC_PROGRAMA	NIVEL_ESTUDIANTE	MATE	APROBADO
count	1670	1670	1670	1670	1670	1670	1670	1670
unique	812	1	5	14	60	12	9	1
top	E10583	202161	2	Facultad Arquitectónica y Artes	Derecho	02	3	N
freq	15	1670	694	380	321	387	473	1670

A partir de estas estadísticas se puede observar lo siguiente:

- Se puede observar que es poco probable que existan valores nulos, ya que todas las variables tienen la misma cantidad de registros.
- Se puede observar que hay 812 estudiantes reprobados, lo que indica una alta probabilidad de que gran parte de estos estudiantes hayan reprobado más de una asignatura durante este periodo.
- El estudiante con mayor número de asignaturas reprobadas es el E11746, con un total de 15 registros de materias reprobadas, sin embargo, es necesario realizar un análisis más detallado, ya que podría existir la posibilidad de registros duplicados o datos atípicos.
- Todos los datos pertenecen únicamente al periodo actual.

- Se presentan 5 tipos de planes de estudio entre los estudiantes.
- Se identifican 14 facultades, y se puede notar que, inicialmente, la Facultad de Arquitectura, Diseño y Artes registra la mayor cantidad de reprobados en este periodo, con 380 casos.
- Se presentan 60 carreras diferentes, siendo la carrera de Derecho la que muestra la mayor cantidad de reprobados, esto sugiere que la facultad no es el único factor determinante para analizar el rendimiento académico, ya que, a pesar de que la Facultad de Arquitectura, Diseño y Artes tiene la mayor cantidad de reprobados en general, la carrera de Derecho supera a todas las carreras individuales de esa facultad en términos de estudiantes reprobados.
- Se registran 12 niveles de estudiantes diferentes.
- Se pueden observar 9 áreas de asignaturas distintas, lo cual es consistente con los registros reportados por la universidad en su sistema Banner.
- Todos los registros corresponden a estudiantes que reprobaron asignaturas.

A continuación, se visualizará las estadísticas descriptivas del segundo periodo del reporte:

Ilustración 157

Estadísticas descriptivas del segundo periodo del reporte de Asignaturas Perdidas

```
categoric_perdidas[1].describe(include="object")
```

	IDD	PERIODO	PLAN_E	DESC_FACULTAD	DESC_PROGRAMA	NIVEL_ESTUDIANTE	MATE	APROBADO
count	2104	2104	2104	2104	2104	2104	2104	2104
unique	1118	1	6	13	55	12	10	1
top	E10583	202166	1	Facultad Arquitectónica y Artes	Derecho	01	3	N
freq	18	2104	1016	424	383	496	567	2104

A partir del despliegue de las estadísticas descriptivas del segundo periodo se pudo observar los siguientes aspectos notorios:

- Se puede observar que el conjunto de datos contiene 2,104 registros en total, que resulta superior al periodo anterior.
- Al igual que en el ejemplo anterior, no hay valores nulos en ninguna de las variables, ya que todas tienen la misma cantidad de registros.
- Se identifican 1,118 estudiantes reprobados en este conjunto de datos, por lo que es visible que hubo una mayor cantidad de estudiantes reprobados de un periodo a otro.
- El estudiante con el mayor número de asignaturas reprobadas es el mismo que en el ejemplo anterior, el estudiante E11746, con un total de 18 registros de materias reprobadas; , por lo que, se puede interpretar que el estudiante ha continuado teniendo dificultades académicas en el periodo actual, sin embargo, también podría referirse a que estos datos son acumulativos, es decir, no solo reportan los estudiantes reprobados en el periodo, sino, guardan el registro y lo suman al reporte de los siguientes periodos.

- Todos los registros corresponden únicamente al periodo actual.
- Se presentan 6 tipos de planes de estudio entre los estudiantes, un incremento respecto al conjunto de datos anterior.
- Se identifican 13 facultades distintas, y nuevamente se observa que la Facultad de Arquitectura, Diseño y Artes tiene la mayor cantidad de reprobados, con 424 casos en este periodo.
- En este periodo, se registran 55 carreras diferentes, siendo Derecho la carrera con la mayor cantidad de reprobados nuevamente, superando a todas las carreras individuales de la Facultad de Arquitectura, Diseño y Artes en términos de estudiantes reprobados.
- Se registran 12 niveles de estudiantes diferentes, lo cual es consistente con el conjunto de datos anterior.
- Se pueden observar 10 áreas de asignaturas distintas, un incremento respecto al conjunto de datos anterior, lo cual es consistente con los registros reportados por la universidad en su sistema Banner.
- Al igual que en el conjunto de datos anterior, todos los registros corresponden a estudiantes que reprobaron asignaturas.

En tercera instancia, se mostrará las estadísticas descriptivas del tercer periodo del reporte de asignaturas perdidas:

Ilustración 158

Estadísticas descriptivas del tercer periodo del reporte de Asignaturas Perdidas

```
categoric_perdidas[2].describe(include="object")
```

	IDD	PERIODO	PLAN_E	DESC_FACULTAD	DESC_PROGRAMA	NIVEL_ESTUDIANTE	MATE	APROBADO
count	3069	3069	3069	3069	3069	3069	3069	3069
unique	1544	1	7	13	57	12	11	1
top	E10583	202261	1	Facultad Ingeniería	Ingeniería Civil	01	3	N
freq	21	3069	1547	864	679	548	722	3069

A partir de las presentes estadísticas, es posible destacar lo siguiente:

- El conjunto de datos contiene 3,069 registros en total, lo que indica un aumento en comparación con los periodos anteriores.
- Al igual que en los ejemplos anteriores, no se observan valores nulos en ninguna de las variables, ya que todas tienen la misma cantidad de registros.
- Se identifican 1,544 estudiantes reprobados en este conjunto de datos, lo que indica un incremento significativo en comparación con los periodos anteriores.
- El estudiante con el mayor número de asignaturas reprobadas es nuevamente el estudiante E11746, con un total de 21 registros de materias reprobadas; por lo que, se puede interpretar que el estudiante ha continuado enfrentando dificultades académicas en el periodo actual, y es posible que estos datos sean acumulativos, sin embargo, de ser acumulativo los registros de este reporte, igualmente presenta que el estudiante reprobó 3 asignaturas este periodo.
- Todos los registros corresponden únicamente al periodo actual.

- Se presentan 7 tipos de planes de estudio entre los estudiantes, lo cual indica una mayor diversidad en comparación con los periodos anteriores.
- Se identifican 13 facultades distintas, y se puede observar que la Facultad de Ingeniería registra la mayor cantidad de reprobados en este periodo, con 864 casos.
- En este periodo, se registran 57 carreras diferentes, siendo Ingeniería Civil la carrera con la mayor cantidad de reprobados, en este caso correspondiendo a la mayor cantidad de probados de la facultad de Ingeniería.
- Se registran 12 niveles de estudiantes diferentes, lo cual es consistente con los conjuntos de datos anteriores.
- Se pueden observar 11 áreas de asignaturas distintas, un incremento respecto a los conjuntos de datos anteriores, lo cual es consistente con los registros reportados por la universidad en su sistema Banner.
- Al igual que en los conjuntos de datos anteriores, todos los registros corresponden a estudiantes que reprobaron asignaturas.

En último término, se exhibirán las estadísticas descriptivas correspondientes al cuarto y último periodo del reporte de asignaturas perdidas:

Ilustración 159

Estadísticas descriptivas del cuarto periodo del reporte de Asignaturas Perdidas

```
categoric_perdidas[3].describe(include="object")
```

	IDD	PERIODO	PLAN_E	DESC_FACULTAD	DESC_PROGRAMA	NIVEL_ESTUDIANTE	MATE	APROBADO
count	3141	3141	3141	3141	3141	3141	3141	3141
unique	1658	1	7	13	51	11	11	1
top	E3110	202266	1	Facultad Ingeniería	Derecho	01	3	N
freq	12	3141	1791	713	546	815	743	3141

A partir de las estadísticas descriptivas del cuarto periodo, se puede destacar lo siguiente:

- El conjunto de datos contiene 3,141 registros en total, lo que indica un aumento en comparación con los periodos anteriores.
- Al igual que en los ejemplos anteriores, no se observan valores nulos en ninguna de las variables, ya que todas tienen la misma cantidad de registros.
- Se identifican 1,658 estudiantes reprobados en este conjunto de datos, lo que indica una cantidad considerable de estudiantes con bajo rendimiento académico.
- El estudiante con el mayor número de asignaturas reprobadas es el estudiante E4386, con un total de 12 registros de materias reprobadas, por lo que, se puede interpretar que el estudiante ha tenido dificultades académicas en el periodo actual, y no se observan datos acumulativos como en los ejemplos anteriores debido a que el estudiante E11746 ya no se encuentra como el estudiante con mayor cantidad de asignaturas reprobadas, sin embargo, es necesario un mayor análisis.
- Todos los registros corresponden únicamente al periodo actual.

- Se presentan 7 tipos de planes de estudio entre los estudiantes, lo cual es consistente con los conjuntos de datos anteriores.
- Se identifican 13 facultades distintas, y se puede observar que la Facultad de Ingeniería registra la mayor cantidad de reprobados en este periodo, con 713 casos.
- En este periodo, se registran 51 carreras diferentes, siendo Derecho la carrera con la mayor cantidad de reprobados, superando a todas las carreras individuales de la Facultad de Ingeniería en términos de estudiantes reprobados.
- Se registran 11 niveles de estudiantes diferentes, lo cual indica una variabilidad en los niveles educativos de los estudiantes reprobados.
- Se pueden observar 11 áreas de asignaturas distintas, lo cual es consistente con los registros reportados por la universidad en su sistema Banner.
- Al igual que en los conjuntos de datos anteriores, todos los registros corresponden a estudiantes que reprobaron asignaturas.

Finalmente se ha finalizado la exploración de las variables categóricas del reporte de asignaturas perdidas con el uso de la técnica de estadísticas descriptivas, de estas, en síntesis, se logró rescatar las siguientes conclusiones:

- El análisis de las estadísticas descriptivas revela que existe una tendencia de reprobación en los periodos estudiados, con un número considerable de estudiantes reprobados en cada conjunto de datos.
- Se observa que un estudiante en particular, identificado como E11746, es recurrente en los conjuntos de datos como el estudiante con el mayor número de asignaturas

reprobadas, por lo que, se puede interpretar que este estudiante enfrenta dificultades académicas persistentes y podría requerir intervención o apoyo adicional, sin embargo, también podría deberse a que el reporte de estudiantes reprobados posea datos acumulativos.

- Las facultades y carreras muestran diferencias en cuanto a la cantidad de estudiantes reprobados. La Facultad de Arquitectura, Diseño y Artes registró la mayor cantidad de reprobados en los primeros conjuntos de datos, mientras que la Facultad de Ingeniería lidera en los conjuntos de datos posteriores. Sin embargo, a pesar de esto, la carrera que más reprobados ha presentado es la de Derecho.
- Existe una diversidad en los planes de estudio, niveles de estudiantes y áreas de asignaturas registradas en cada conjunto de datos, lo que refleja la variedad de programas académicos ofrecidos por la universidad y resalta la importancia de considerar estos aspectos al analizar el rendimiento académico.
- Es importante tener en cuenta que los conjuntos de datos presentan datos acumulativos en algunos casos, lo que significa que los registros de asignaturas reprobadas pueden incluir periodos anteriores; es decir, esto podría afectar la interpretación de los resultados y requiere un análisis más detallado para comprender completamente la situación académica de los estudiantes.

Ahora, a causa de la culminación del análisis por estadísticas descriptivas, a partir de este momento se realizará una exploración de las variables categóricas del mismo reporte de asignaturas perdidas, haciendo uso de un diagrama de pastel o de barras dependiendo exclusivamente de la variable, no obstante, cabe remarcar que las razones por las que se usará

una técnica u otra se explicará en conjunto con la visualización de sus respectivas ilustraciones, sin embargo, no se profundizará en las explicaciones con la visualizaciones de sus razones, debido a que, las razones que serán descritas por cada variable serán fácilmente comprobables al momento de visualizarse el gráfico de la técnica seleccionada.

Antes de comenzar el análisis, se presentará un código que ha servido para la mejor presentación y visualización de los gráficos de cada variable, el código es:

Ilustración 160

Función necesaria para mejorar la visualización de los gráficos

```
def get_periodo(id_dataframe):  
    """  
    Devuelve el nombre del periodo dado el id del dataframe de la lista de dataframes de los reportes  
    """  
    nombre, data_id = id_dataframe.split("0")  
    if data_id == "1": return "Periodo 2021-01"  
    elif data_id == "2": return "Periodo 2021-02"  
    elif data_id == "3": return "Periodo 2022-01"  
    elif data_id == "4": return "Periodo 2022-02"
```

Ahora, para comenzar el análisis se iniciará con la exploración de la variable PLAN_E que se refiere al plan de estudios del estudiante, esta variable, debido a que dispone de valores muy desproporcionados entre sus categorías, se ha tomado la decisión de utilizar la técnica de gráfico de barras.

Ilustración 161

Código que genera el gráfico de la variable PLAN_E del reporte de Asignaturas Perdidas

```
# Se especifica el número de filas y columnas para los subplots
nrows = 2
ncols = 2

# Se crea la figura y los ejes de los subplots con el tamaño especificado
fig, axes = plt.subplots(nrows=nrows, ncols=ncols, figsize=(15, 10))

# Se establece el nombre de la columna de interés
columna_interes = 'PLAN_E'

# Se itera a través de los dataframes y se crea un gráfico de barras para cada uno
for idx, (perdida, ax) in enumerate(zip(perdidas, axes.flat)):

    # Se utiliza seaborn para generar un gráfico de barras de la columna de interés del dataframe actual
    sns.countplot(x=columna_interes, data=perdida, ax=ax)

    # Se establecen las etiquetas de los ejes x, y y el título para cada subplot
    ax.set_xlabel(f'{columna_interes}')
    ax.set_ylabel('Cantidad')
    ax.set_title(f'Gráfico de barras de {columna_interes} ({get_periodo(perdida.name)})')

# Se ajusta el espacio entre los subplots para que no se superpongan
plt.subplots_adjust(wspace=0.2, hspace=0.4)

# Se muestra la figura con los subplots
plt.show()
```

Ilustración 162

Gráficos de barras del primer y segundo periodo de la variable PLAN_E del reporte de Asignaturas Perdidas

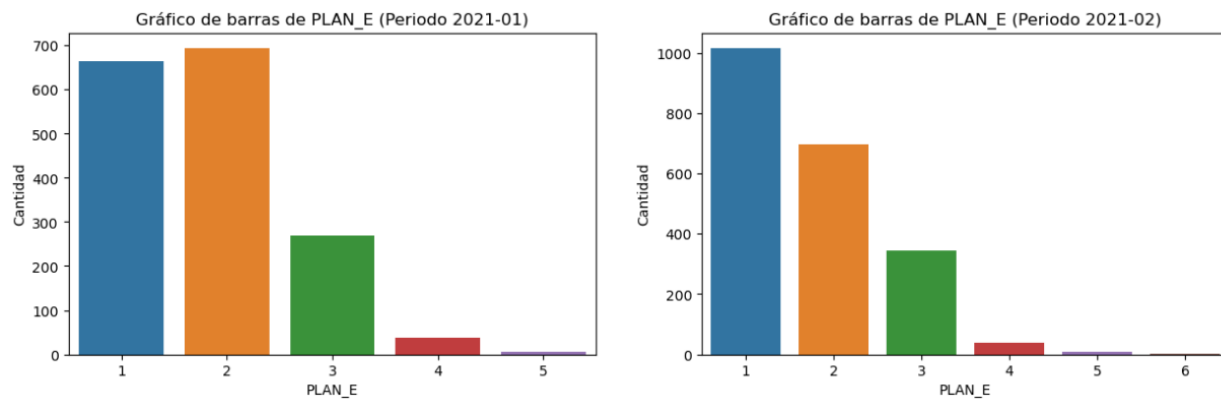
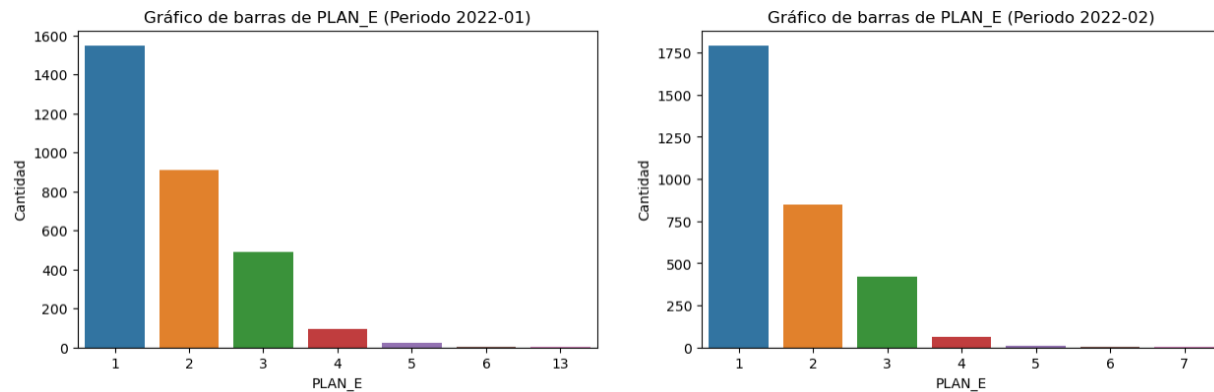


Ilustración 163

Gráficos de barras del segundo y tercer periodo de la variable PLAN_E del reporte de Asignaturas Perdidas



A partir de los gráficos de barras se puede concluir lo siguiente:

- Es observable que los planes de estudio 1 y 2 son los que más reprobados presentan.
- Es visible que los planes 5, 6, 7 y 13 no presentan una gran cantidad de estudiantes reprobados en comparación a los demás planes.
- Hasta el momento, es posible que la variable PLAN_E tiene posibilidad de ser un factor importante para analizar si un estudiante reprobará o no una asignatura, debido a que los datos de las asignaturas perdidas incluyen solo aquellos estudiantes que han reprobado.

En segundo lugar, se analizará la variable DESC_FACULTAD, la cual se refiere a la facultad a la que pertenecen los estudiantes hallados en el reporte de asignaturas perdidas, para esta ocasión, de igual manera se utilizarán gráficos de barras, principalmente debido a la desproporción de los datos.

Ilustración 164

Código que genera el gráfico de la variable DESC_FACULTAD del reporte de Asignaturas Perdidas

```
# Ordenamos las categorías alfabéticamente
orden_categorias = sorted(set(perdidas[0]["DESC_FACULTAD"].unique()))

# Se especifica el número de filas y columnas para los subplots
nrows = 2
ncols = 2

# Se crea la figura y los ejes de los subplots con el tamaño especificado
fig, axes = plt.subplots(nrows=nrows, ncols=ncols, figsize=(15, 25))

# Se establece el nombre de la columna de interés
columna_interes = 'DESC_FACULTAD'

# Se itera a través de los dataframes y se crea un gráfico de barras para cada uno
for idx, (perdida, ax) in enumerate(zip(perdidas, axes.flat)):
    # Se utiliza seaborn para generar un gráfico de barras de la columna de interés del dataframe actual
    sns.countplot(x=columna_interes, data=perdida, ax=ax, order=orden_categorias)
    # Se establecen las etiquetas de los ejes x, y el título para cada subplot
    ax.set_xlabel(columna_interes)
    ax.set_xticklabels(ax.get_xticklabels(), rotation=90)
    ax.set_ylabel('Cantidad')
    ax.set_title(f'Gráfico de barras de {columna_interes} ({perdida.name})')

# Se ajusta el espacio entre los subplots para que no se superpongan
plt.subplots_adjust(wspace=0.4, hspace=0.4)

# Se muestra la figura con los subplots
plt.show()
```

Ilustración 165

Gráficos de barras del primer y segundo periodo de la variable DESC_FACULTAD del reporte de Asignaturas Perdidas

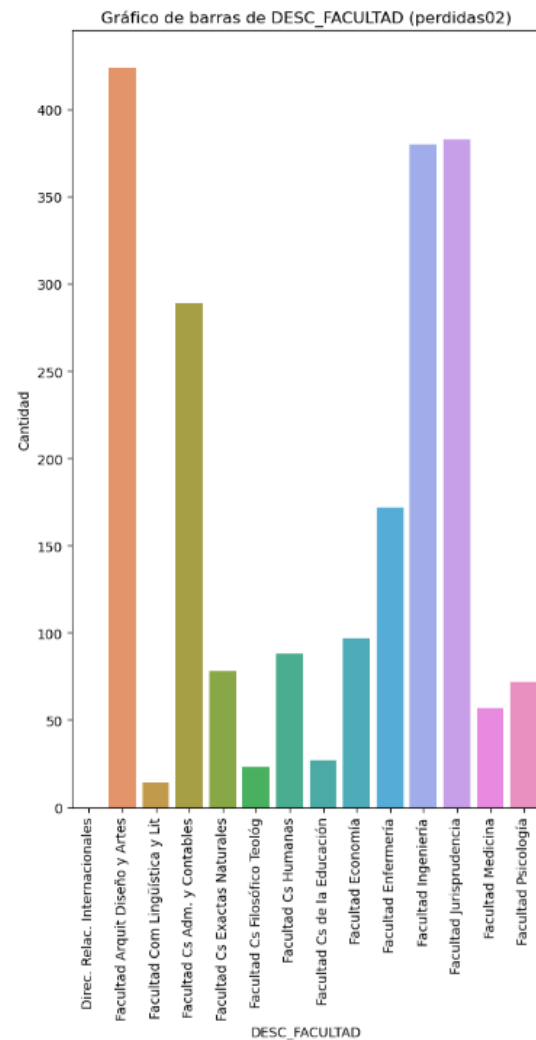
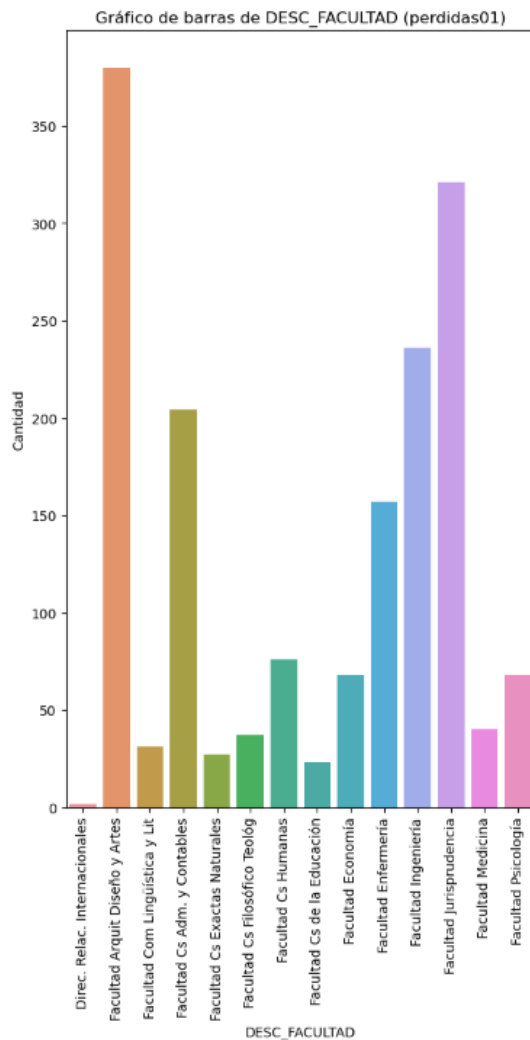
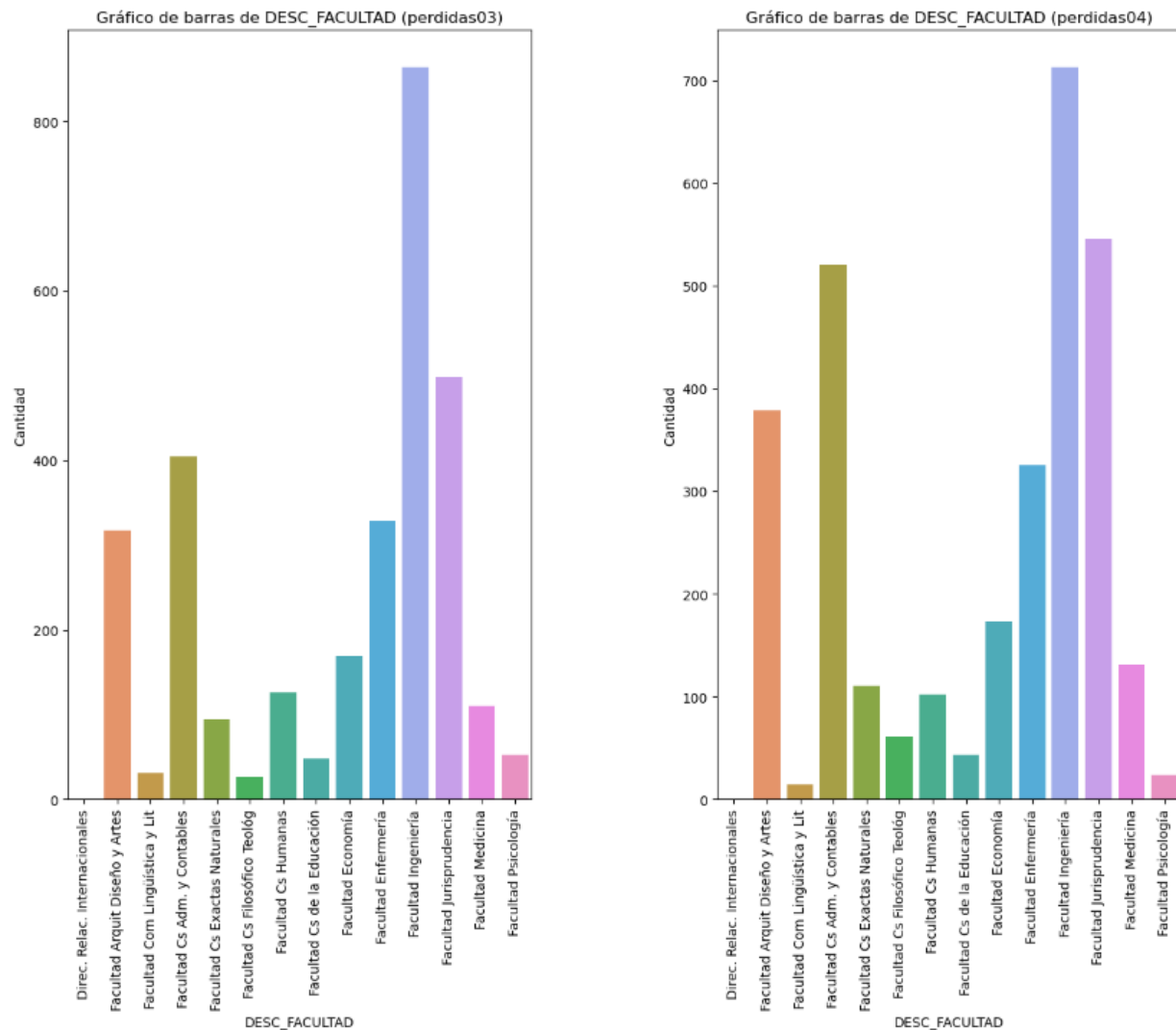


Ilustración 166

Gráficos de barras del segundo y tercer periodo de la variable DESC_FACULTAD del reporte de Asignaturas Perdidas



A partir del análisis realizado sobre la variable DESC_FACULTAD, se concluyó que:

- Se puede observar que las facultades con mayor cantidad de reprobados son:
 - Facultad de Arquitectura, Diseño y Artes
 - Facultad de Ingeniería
 - Facultad de Jurisprudencia

- Facultad de Ciencias Administrativas y Contables
- Facultad de Enfermería
- Durante todos los periodos este patrón se repite, donde las facultades mencionadas siempre corresponden a la mayor cantidad de reprobados. Esto podría indicar una cierta relación en la que la facultad aumente o disminuya la probabilidad de que un estudiante repruebe una asignatura.

En tercer lugar, se realizará el análisis de la variable DESC_PROGRAMA, la cual se refiere a la carrera que pertenecen los estudiantes hallados en el reporte de asignaturas perdidas; en esta ocasión, considerando que existen un gran número de carreras universitarias, se tomó la decisión de utilizar la técnica de gráficos de barras.

Ilustración 167

Código que genera el gráfico de la variable DESC_PROGRAMA del reporte de Asignaturas Perdidas

```
# Ordenamos las categorías alfabéticamente
orden_categorias = sorted(set(perdidas[0]["DESC_PROGRAMA"].unique()))

# Se especifica el número de filas y columnas para los subplots
nrows = 4
ncols = 1

# Se crea la figura y los ejes de los subplots con el tamaño especificado
fig, axes = plt.subplots(nrows=nrows, ncols=ncols, figsize=(15, 30))

# Se establece el nombre de la columna de interés
columna_interes = 'DESC_PROGRAMA'

# Se itera a través de los dataframes y se crea un gráfico de barras para cada uno
for idx, (perdida, ax) in enumerate(zip(perdidas, axes.flat)):
    # Se utiliza seaborn para generar un gráfico de barras de la columna de interés del dataframe actual
    sns.countplot(x=columna_interes, data=perdida, ax=ax, order=orden_categorias)
    # Se establecen las etiquetas de los ejes x, y y el título para cada subplot
    ax.set_xlabel(f"{columna_interes}")
    ax.set_xticklabels(ax.get_xticklabels(), rotation=90)
    ax.set_ylabel('Cantidad')
    ax.set_title(f'Gráfico de barras de {columna_interes} ({get_periodo(perdida.name)})')

# Se ajusta el espacio entre los subplots para que no se superpongan
plt.subplots_adjust(wspace=0, hspace=1)

# Se muestra la figura con los subplots
plt.show()
```

Ilustración 168

Gráficos de barras del primer periodo de la variable DESC_PROGRAMA del reporte de Asignaturas Perdidas

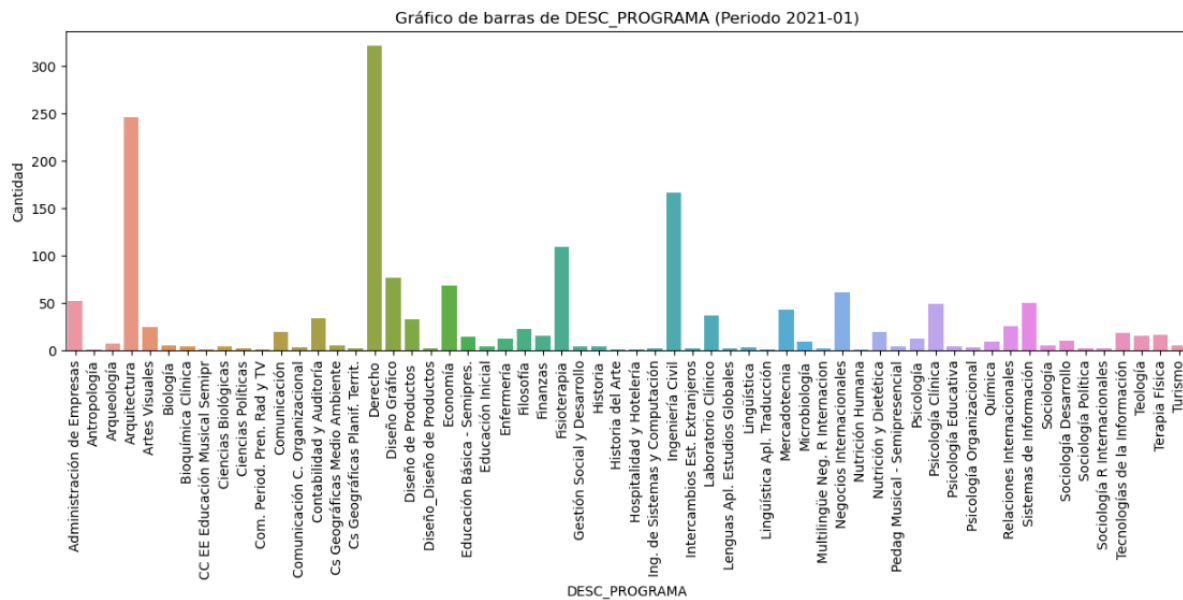


Ilustración 169

Gráficos de barras del segundo periodo de la variable DESC_PROGRAMA del reporte de Asignaturas Perdidas

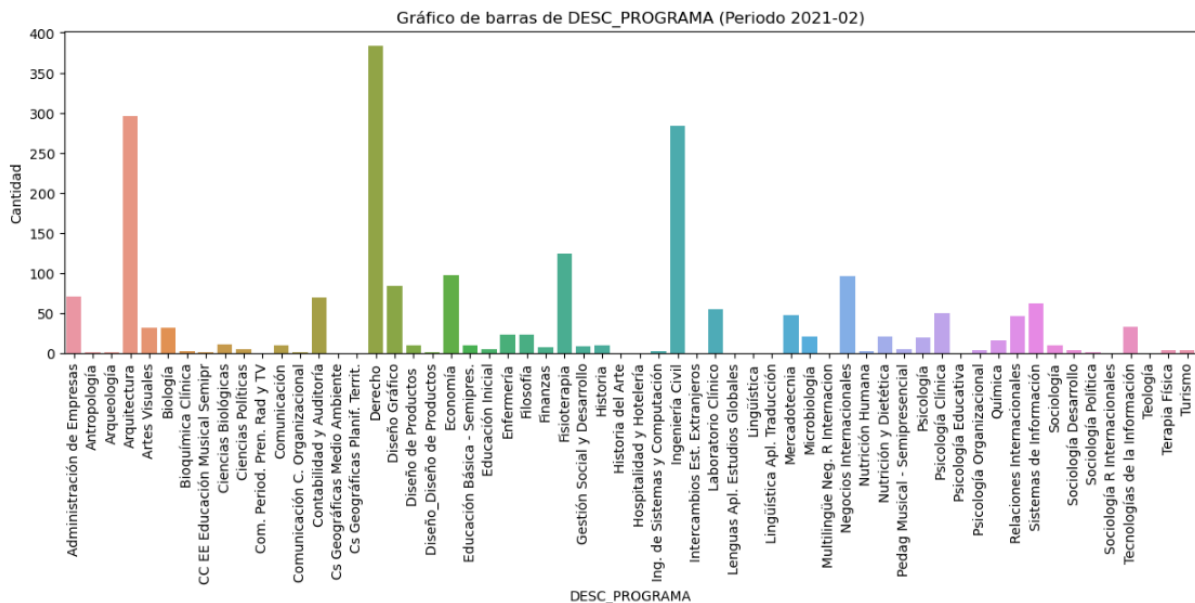


Ilustración 170

Gráficos de barras del tercer periodo de la variable DESC_PROGRAMA del reporte de Asignaturas Perdidas

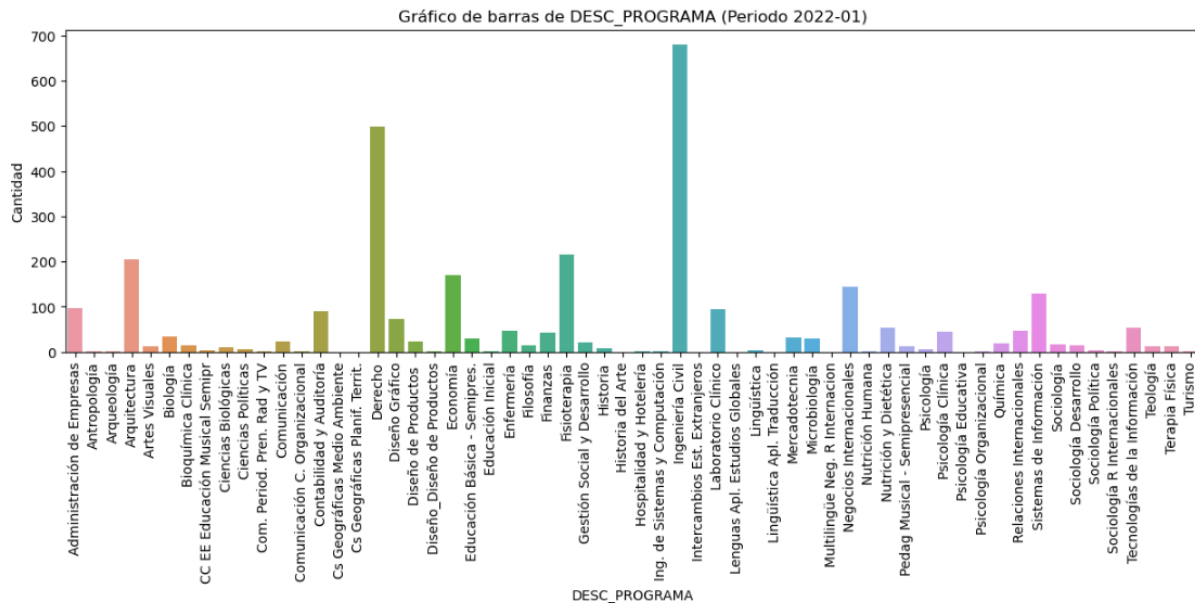
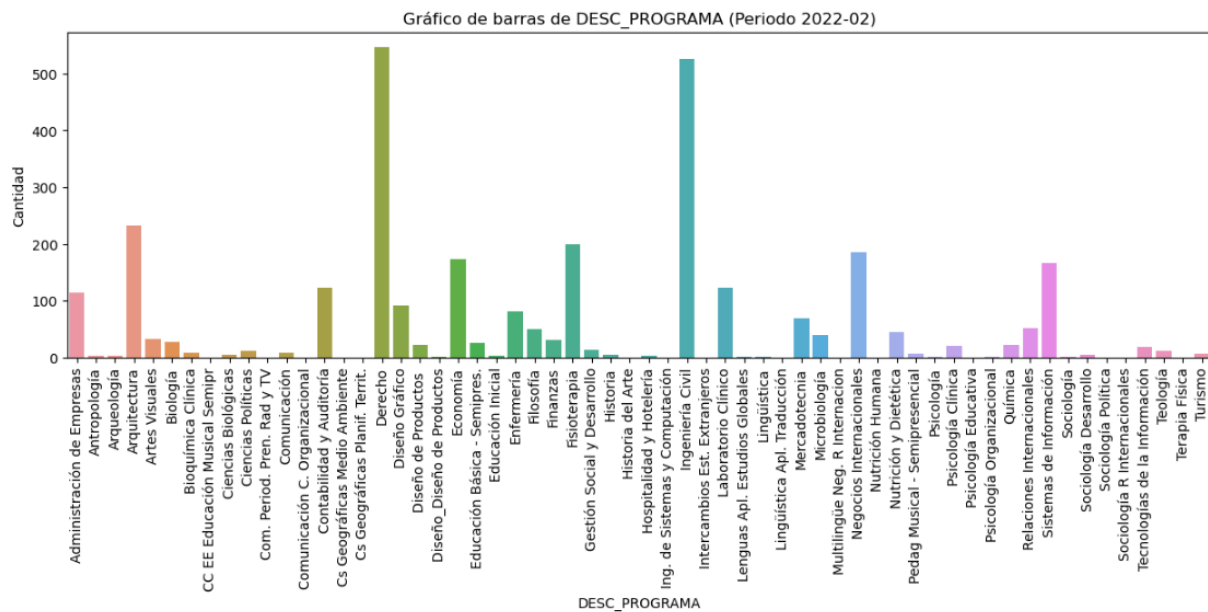


Ilustración 171

Gráficos de barras del cuarto periodo de la variable DESC_PROGRAMA del reporte de Asignaturas Perdidas



A partir del análisis realizado para la variable DESC_PROGRAMA, se puede concluir lo siguiente:

- Se puede observar como las carreras que presentan la mayor cantidad de asignaturas reprobadas pertenecen a:
 - Derecho
 - Arquitectura
 - Ingeniería civil
 - Fisioterapia
- Sin embargo, también es visible como surge una tendencia en la que carreras como las de Sistemas de Información aumentan progresivamente su cantidad de asignaturas reprobadas.
- En general, se observa un comportamiento similar a lo largo de los periodos dentro de una misma carrera, pero no entre diferentes carreras, lo que podría indicar que la carrera en la que un estudiante se encuentra podría tener un impacto en su rendimiento académico.

En cuarto lugar, se realizará el análisis de los niveles académicos de los estudiantes universitarios, para esta ocasión de igual manera se considerará el uso de gráficos de barras, debido a que, existe una gran disparidad de datos y la cantidad de sus categorías haría que resultará inviable un análisis sobre un gráfico de pastel.

Ilustración 172

Código que genera el gráfico de la variable NIVEL_ESTUDIANTE del reporte de Asignaturas Perdidas

```
# Ordenamos las categorías alfabéticamente
orden_categorias = sorted(set(perdidas[0]["NIVEL_ESTUDIANTE"].unique()))

# Número de filas y columnas para los subplots
nrows = 2
ncols = 2

# Crear la figura y los ejes para los subplots
fig, axes = plt.subplots(nrows=nrows, ncols=ncols, figsize=(15, 15))

# Nombre de la columna de interés
columna_interes = 'NIVEL_ESTUDIANTE'

# Iterar a través de los dataframes y crear un gráfico para cada uno
for idx, (perdida, ax) in enumerate(zip(perdidas, axes.flat)):
    # Se utiliza seaborn para generar un gráfico de barras de la columna de interés del dataframe actual
    sns.countplot(x=columna_interes, data=perdida, ax=ax, order=orden_categorias)
    # Se establecen las etiquetas de los ejes x, y y el título para cada subplot
    ax.set_xlabel(f"{columna_interes}")
    ax.set_ylabel('Cantidad')
    ax.set_title(f'Gráfico de barras de {columna_interes} ({get_periodo(perdida.name)})')

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.4)

# Mostrar la figura
plt.show()
```

Ilustración 173

Gráficos de barras del primer y segundo periodo de la variable NIVEL_ESTUDIANTE del reporte de Asignaturas Perdidas

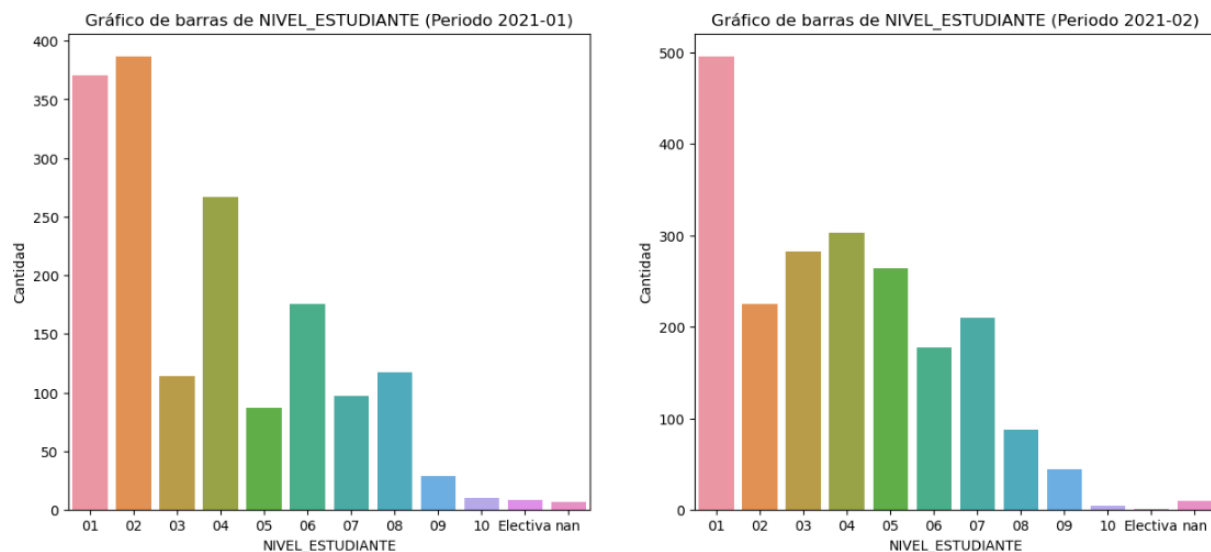
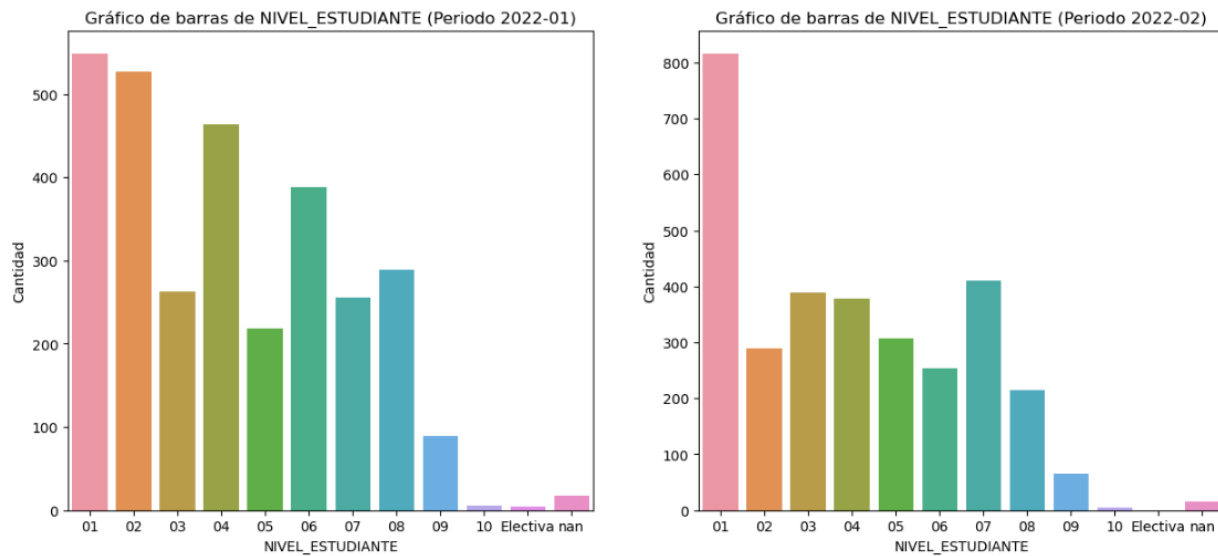


Ilustración 174

Gráficos de barras del tercer y cuarto periodo de la variable NIVEL_ESTUDIANTE del reporte de Asignaturas Perdidas



Los análisis sobre los gráficos de barras de la variable NIVEL_ESTUDIANTE permiten las siguientes conclusiones:

- Dado a que los estudiantes registrados en el reporte de asignaturas perdidas se deben a que reprobaron determinadas asignaturas, se puede concluir que aparentemente no podría ser un factor determinante el nivel de un estudiante para reprobado, sin embargo, también es visible que aquellos estudiantes en un nivel de noveno o decimo semestre tiene menor cantidad de estudiante, lo que podría sugerir dos opciones:
 - Existe menor probabilidad de reprobado si el estudiante se encuentra en noveno o decimo semestre.
 - Existe una menor cantidad de registros de estudiantes en noveno o decimo semestre.
- Es observable la presencia de valores faltantes.

En quinto lugar, se realizará el análisis exploratorio de la variable MATE, es decir, del área de asignaturas del reporte de asignaturas perdidas, sin embargo, es necesario recordar que esta variable permitirá identificar a las asignaturas que estudiantes determinados cursarán o están pensando en cursar. Por otra parte, para esta ocasión, semejante a las demás hasta el momento, se utilizará la técnica del gráfico de barras debido a la desproporción de los datos, ya que, hacen inviable cualquier análisis sobre un posible gráfico de barras sobre sus datos.

Ilustración 175

Código que genera el gráfico de la variable MATE del reporte de Asignaturas Perdidas

```
# Ordenamos las categorías alfabéticamente
orden_categorias = sorted(set(perdidas[0]["MATE"].unique()))

# Número de filas y columnas para los subplots
nrows = 2
ncols = 2

# Crear la figura y los ejes para los subplots
fig, axes = plt.subplots(nrows=nrows, ncols=ncols, figsize=(15, 15))

# Nombre de la columna de interés
columna_interes = 'MATE'

# Iterar a través de los dataframes y crear un gráfico para cada uno
for idx, (perdida, ax) in enumerate(zip(perdidas, axes.flat)):
    # Se utiliza seaborn para generar un gráfico de barras de la columna de interés del dataframe actual
    sns.countplot(x=columna_interes, data=perdida, ax=ax, order=orden_categorias)
    # Se establecen las etiquetas de los ejes x, y y el título para cada subplot
    ax.set_xlabel(f"{columna_interes}")
    ax.set_ylabel('Cantidad')
    ax.set_title(f'Gráfico de barras de {columna_interes} ({get_periodo(perdida.name)})')

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.2)

# Mostrar la figura
plt.show()
```

Ilustración 176

Gráficos de barras del primer y segundo periodo de la variable MATE del reporte de Asignaturas Perdidas

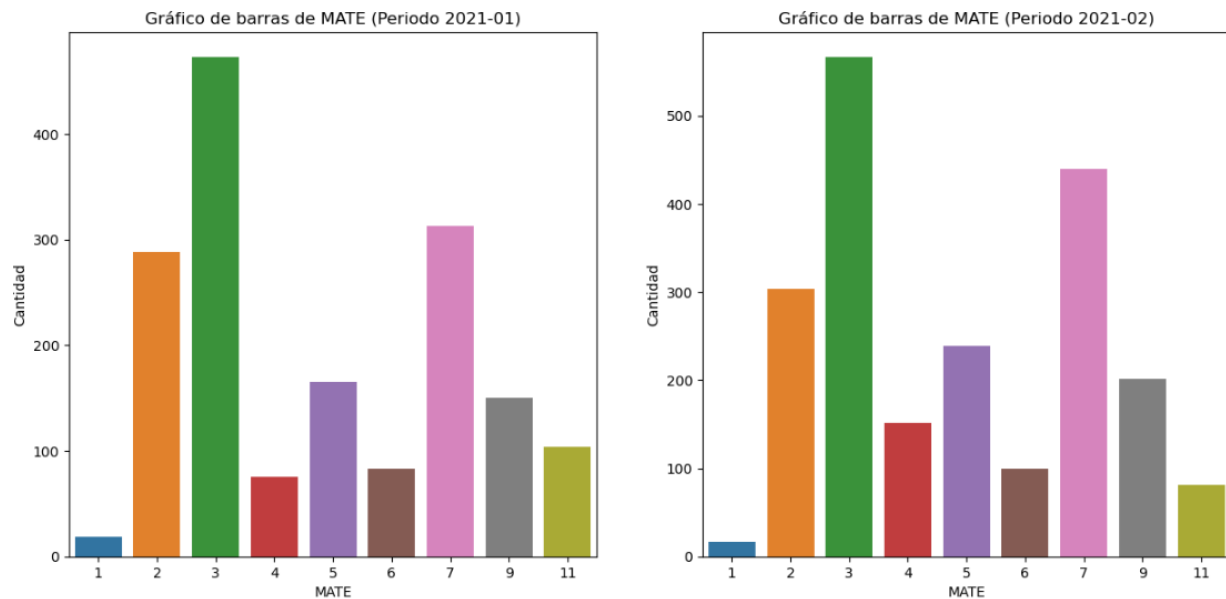
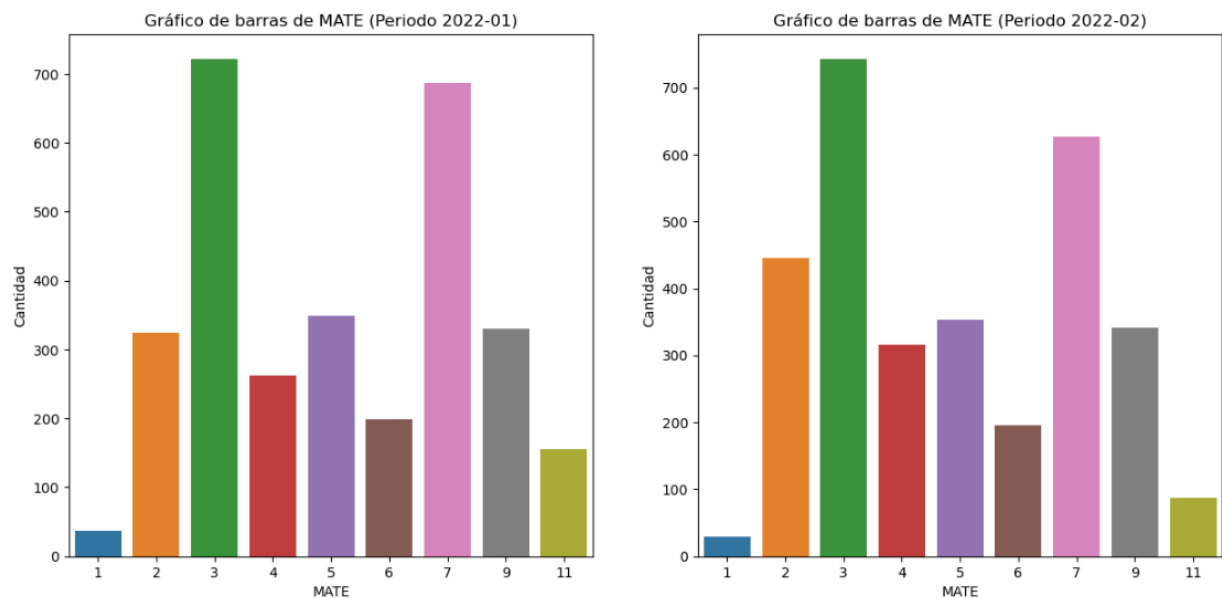


Ilustración 177

Gráficos de barras del tercer y cuarto periodo de la variable MATE del reporte de Asignaturas Perdidas



A partir del análisis realizado de la variable MATE, es decir, del área de la asignatura que el estudiante reprobó, se puede concluir que existe un patrón de comportamiento similar a través

de los períodos con respecto a las áreas de las asignaturas reprobadas, es decir, es posible que exista una relación entre si un estudiante aprobará o no una asignatura, si esta pertenece a un área de asignatura específica. Adicionalmente, se recuerda al distinguido lector que la variable MATE, y sus respectivas representaciones en los distintos reportes, se prevé que sean la manera de identificar a la asignatura que un estudiante va a cursar.

En sexto y último lugar, se realizará un análisis sobre la variable APROBADO, la cual indica el estado de aprobación de los estudiantes, para esta ocasión, a diferencia de todos los demás análisis realizados hasta la fecha, se utilizará el método de gráfico de pastel debido a que sus datos permiten un análisis prudente de estos con esta técnica.

Ilustración 178

Código que genera el gráfico de la variable APROBADO del reporte de Asignaturas Perdidas

```
# Se ordenan las categorías alfabéticamente
orden_categorias = sorted(set(perdidas[0]["APROBADO"].unique()))

# Se define el número de filas y columnas para los subplots
nrows = 2
ncols = 2

# Se crea la figura y los ejes para los subplots
fig, axes = plt.subplots(nrows=nrows, ncols=ncols, figsize=(15, 15))

# Se especifica el nombre de la columna de interés
columna_interes = 'APROBADO'

# Colores personalizados para el gráfico de pie
colores = ['#D5303E']

# Se itera a través de los dataframes y se crea un gráfico de pie para cada uno
for idx, (perdida, ax) in enumerate(zip(perdidas, axes.flat)):
    # Se obtiene la cantidad de cada categoría en la columna de interés
    counts = perdida[columna_interes].value_counts()
    # Se obtienen las etiquetas y los tamaños para el gráfico de pie
    labels = counts.index
    sizes = counts.values
    # Se crea el gráfico de pie en los ejes actuales
    ax.pie(sizes, labels=labels, autopct='%1.1f%%', startangle=90, colors=colores)
    # Se establece el título para cada subplot
    ax.set_title(f'Gráfico de pastel de {columna_interes} ({get_periodo(perdida.name)})')

# Se ajusta el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.2)

# Se muestra la figura
plt.show()
```

Ilustración 179

Gráficos de pastel del primer y segundo periodo de la variable APROBADO del reporte de Asignaturas Perdidas

Gráfico de pastel de APROBADO (Periodo 2021-01)

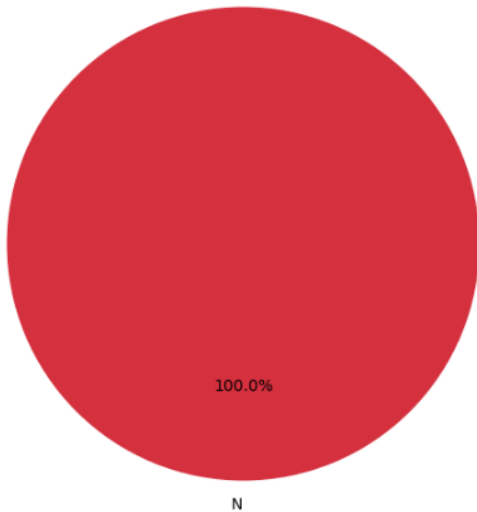


Gráfico de pastel de APROBADO (Periodo 2021-02)

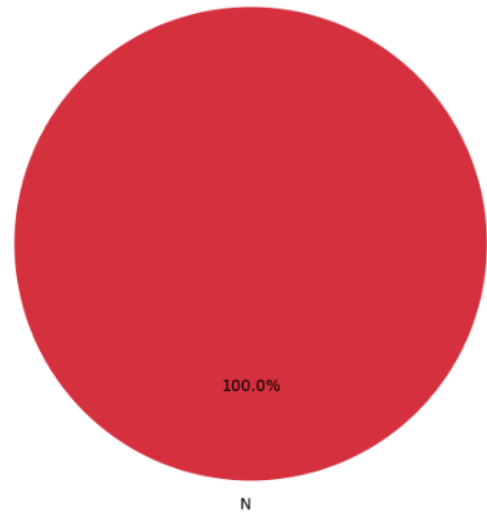


Ilustración 180

Gráficos de pastel del tercer y cuarto periodo de la variable APROBADO del reporte de Asignaturas Perdidas

Gráfico de pastel de APROBADO (Periodo 2022-01)

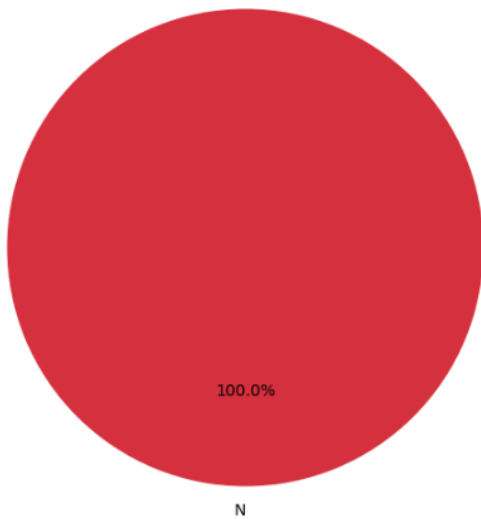
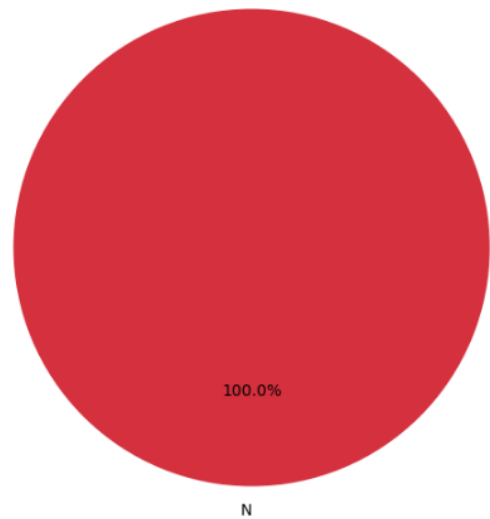


Gráfico de pastel de APROBADO (Periodo 2022-02)



A partir de la visualización y análisis de los gráficos de pastel del estado de aprobación de los estudiantes del reporte de asignaturas perdidas, se puede concretar lo obvio, es decir, se puede concluir que todos los registros de todos los periodos del reporte pertenecen a asignaturas reprobadas.

De manera concluyente, a partir de los análisis realizados sobre todas las variables categóricas del reporte de asignaturas perdidas, se pudieron destacar las siguientes variables de interés:

- PLAN_E
- DESC_FACULTAD
- DESC_PROGRAMA
- MATE

Por otra parte, a pesar de que la variable NIVEL_ESTUDIANTE no presentase una mayor relevancia, es posible que aun pueda serlo dado los análisis, por lo que, es necesario incluirla en los análisis de relaciones posteriores. Sin embargo, por el momento, se continuará con a la exploración de las variables categóricas del reporte de calificaciones por carrera, comenzando propiamente por las estadísticas descriptivas del primer periodo del reporte:

Ilustración 181

Estadísticas Descriptivas del primer periodo del reporte de Calificaciones por Carrera

categoric_calificaciones[0].describe()							
	id_banner_estudiante	codigo_periodo	facultad	id_banner_docente	descripcion_materia	descripcion_programa	nota_final_SHATCKG
count	163810	163810	163810	163810	163810	163810	163810
unique	7587	1	14	713	11	74	63
top	E5814	202161	Facultad Enfermería	D591	Cien.Soc.period.form y Derecho	Derecho	45
freq	44	163810	26232	1132	42509	21984	12645

A partir de las presentes estadísticas descriptivas se pudo observar los siguientes comportamientos:

- El conjunto de datos contiene un total de 163,810 registros.
- Es posible que existan valores nulos en el reporte dada la cantidad de registros que se presentan en algunas variables.
- Se identifican 7,587 estudiantes únicos en el conjunto de datos, lo que quiere decir que estas estudiantes están inscritas en varias asignaturas durante este periodo.
- Todos los registros pertenecen al mismo periodo.
- Se registran 14 facultades diferentes en el conjunto de datos. La Facultad de Enfermería es la que tiene la mayor cantidad de registros, con 26,232 casos.
- Existen 713 docentes diferentes identificados en el conjunto de datos, lo que sugiere que hay varios estudiantes asignados a un mismo profesor.
- Se registran 11 áreas de asignaturas diferentes, indicando la variedad de asignaturas en los registros debido a que corresponden a todas las áreas posibles de asignaturas.
- Se observa un total de 163,791 registros en la variable "secuencial_parcial". La diferencia en el recuento con respecto a las otras variables puede indicar la presencia de valores nulos o registros faltantes en esa variable específica.
- Se puede observar que existen 74 carreras reportadas, este es un valor superior al de las carreras reportadas en el reporte de asignaturas perdidas.
- La carrera en la que más estudiantes se encuentran matriculados es la de Derecho, sin embargo, la facultad con mayor número de estudiantes es la Facultad de Enfermería.
- El área de asignatura más cursado es el de Ciencias sociales, periodismo y derecho.

En segundo término, se presentarán las estadísticas descriptivas del segundo periodo del reporte:

Ilustración 182

Estadísticas Descriptivas del segundo periodo del reporte de Calificaciones por Carrera

```
categoric_calificaciones[1].describe(include="object")
```

	id_banner_estudiante	codigo_perodo	facultad	id_banner_docente	descripcion_materia	descripcion_programa	nota_final_SHATCKG
count	158818	158818	158818	158818	158818	158818	158818
unique	7629	1	14	738	11	72	65
top	E8581	202166	Facultad Enfermería	D596	Cien.Soc.period.form y Derecho	Derecho	50
freq	44	158818	23076	892	39474	22024	11220

A partir de las estadísticas demostradas, se logró observar lo siguiente:

- El conjunto de datos contiene un total de 158,818 registros.
- No se observan valores nulos en ninguna de las variables, ya que todas tienen la misma cantidad de registros.
- Se identifican 7,629 estudiantes únicos en el conjunto de datos, es decir, aumento el número de estudiante matriculados con respecto al anterior semestre.
- Los registros de datos corresponden a un solo periodo, el actual evaluado.
- Se registran 14 facultades diferentes en el conjunto de datos. La Facultad de Enfermería es la que tiene la mayor cantidad de registros, con 23,076 casos.
- Existen 738 docentes diferentes identificados en el conjunto de datos, es decir, la universidad contrato más docentes para el periodo actual.
- Se registran 11 áreas de asignaturas diferentes, indicando la variedad de asignaturas en los registros debido a que corresponden a todas las áreas posibles de asignaturas.

- Se puede observar que existen 72 carreras reportadas, este es un valor igualmente superior al de las carreras reportadas en el reporte de asignaturas perdidas, además, es visible que dos carreras cesaron de existir.
- La carrera en la que más estudiantes se encuentran matriculados es la de Derecho, sin embargo, la facultad con mayor número de estudiantes es la Facultad de Enfermería igualmente.
- El área de asignatura más cursado es de igual manera el de Ciencias sociales, periodismo y derecho.

Como tercera consideración, se presentarán las estadísticas descriptivas del tercer periodo del reporte de calificaciones por carrera:

Ilustración 183

Estadísticas Descriptivas del tercer periodo del reporte de Calificaciones por Carrera

categoric_calificaciones[2].describe(include="object")							
	id_banner_estudiante	codigo_periodo	facultad	id_banner_docente	descripcion_materia	descripcion_programa	nota_final_SHATCKG
count	150108	150108	150108	150108	150108	150108	150108
unique	7205	1	14	787	11	68	67
top	E10452	202261	Facultad Ingeniería	D427	Cien.Soc.period.form y Derecho	Derecho	50
freq	44	150108	21732	1516	41444	22716	10390

A partir de las estadísticas descriptivas del tercer periodo del reporte, se puede destacar lo siguiente:

- El conjunto de datos contiene un total de 150,108 registros.
- No se observan valores nulos en ninguna de las variables, ya que todas tienen la misma cantidad de registros.

- Se identifican 7,205 estudiantes únicos en el conjunto de datos, lo que implica que hubo una menor cantidad de estudiantes matriculados, esto puede darse a retiros académicos, finalización de la carrera universitaria, etc.
- Los registros de datos corresponden a un solo periodo, el actual evaluado.
- Se registran 14 facultades diferentes en el conjunto de datos. La Facultad de Ingeniería es la que tiene la mayor cantidad de registros, con 21,732 casos.
- Hay 787 docentes diferentes identificados en el conjunto de datos, lo que indica que progresivamente se van contratando nuevos docentes.
- Se registran 11 áreas de asignaturas diferentes, indicando la variedad de asignaturas en los registros debido a que corresponden a todas las áreas posibles de asignaturas.
- Se puede observar que existen 68 carreras reportadas, este es un valor igualmente superior al de las carreras reportadas en el reporte de asignaturas perdidas, además, es visible que una mayor cantidad de carreras cesaron de existir o que estudiantes simplemente no se matricularon en ellas.
- La carrera en la que más estudiantes se encuentran matriculados es la de Derecho, sin embargo, la facultad con mayor número de estudiantes es la Facultad de Ingeniería.
- De todas maneras, el área de asignatura más cursado es de igual manera el de Ciencias sociales, periodismo y derecho.

Finalmente, ha llegado el momento de presentar las estadísticas correspondientes al cuarto y último periodo del reporte:

Ilustración 184

Estadísticas Descriptivas del cuarto periodo del reporte de Calificaciones por Carrera

```
categoric_calificaciones[3].describe(include="object")
```

	id_banner_estudiante	codigo_periodo	facultad	id_banner_docente	descripcion_materia	descripcion_programa	nota_final_SHATCKG
count	142205	142205	142205	142205	142205	142205	142205
unique	6950	1	14	774	11	69	70
top	E10464	202266	Facultad Jurisprudencia	D104	Cien.Soc.period.form y Derecho	Derecho	50
freq	36	142205	21080	1032	37157	23898	8577

A partir de las estadísticas presentadas, se puede observar lo siguiente:

- El conjunto de datos contiene un total de 142,205 registros.
- No se observan valores nulos en ninguna de las variables, ya que todas tienen la misma cantidad de registros.
- Se identifican 6,950 estudiantes únicos en el conjunto de datos, lo que indica que hubo una disminución en la cantidad de estudiantes matriculados; esta reducción puede ser atribuida a diversos factores, como retiros académicos, finalización de la carrera universitaria, entre otros. De todas maneras, en general se observa una tendencia hacia la disminución de estudiantes en la universidad, lo cual puede estar relacionado con el análisis previo realizado sobre las notas finales, donde se evidenció una disminución en el rendimiento académico.
- Los registros de datos corresponden a un solo periodo, el actual evaluado.
- Se registran 14 facultades diferentes en el conjunto de datos. La Facultad de Jurisprudencia es la que tiene la mayor cantidad de registros, con 21,080 casos.
- Existen 774 docentes diferentes identificados en el conjunto de datos, lo que indica que por primera vez se redujo la cantidad de docentes contratados en la universidad. Es probable que se deba a la reducción de la cantidad de estudiantes matriculados.

- Se registran 11 áreas de asignaturas diferentes, indicando la variedad de asignaturas en los registros debido a que corresponden a todas las áreas posibles de asignaturas.
- Se puede observar que existen 69 carreras reportadas, este es un valor igualmente superior al de las carreras reportadas en el reporte de asignaturas perdidas, además, es visible que en esta ocasión una carrera comenzó a volver a tener estudiantes matriculados o se creó una nueva.
- La carrera en la que más estudiantes se encuentran matriculados es la de Derecho, que corresponde con la Facultad de Jurisprudencia que igualmente indica el mayor número de estudiantes registrados.
- De todas maneras, el área de asignatura más cursado es de igual manera el de Ciencias sociales, periodismo y derecho.

Una vez se recopilaron todas las observaciones de cada estadística descriptiva correspondiente a cada uno de los cuatro periodos del reporte de calificaciones por carrera, se pudo concluir lo siguiente de las estadísticas:

- A lo largo de los periodos analizados, se observa cierta consistencia en el tipo de datos recopilados, como que siempre se mantienen las 11 áreas de asignaturas y 14 facultades, aunque cambia la distribución de los estudiantes entre ellas.
- Se evidencia una disminución en la cantidad de estudiantes únicos matriculados a lo largo de los periodos estudiados, por lo que podría ser el resultado de diversos factores como retiros académicos, graduación, pausas académicas, entre otros.
- Se observa que algunas carreras cesan de existir, mientras otras nuevas aparecen o retoman matriculaciones; por otra parte, a pesar de que en algunos periodos se han disminuido las carreras existentes, se observa que el número de docentes ha

aumentado hasta el tercer periodo, para luego disminuir en el cuarto; este último cambio en el patrón podría estar relacionado con el descenso en la matrícula estudiantil.

- Aunque la carrera de Derecho ha sido consistentemente la más popular entre los estudiantes, la facultad con más matrículas ha variado entre Enfermería, Ingeniería y Jurisprudencia. Esta información puede ser útil para la planificación de recursos y estrategias académicas, además, permite la suposición de que las predicciones para estudiantes de estas carreras serán más certeras.
- Finalmente, enfocándose en la disminución progresiva de estudiantes en la universidad, es posible que exista una relación directa con el rendimiento académico que, como se demostró anteriormente, a través de los periodos académicos esta ha ido en descenso.

No obstante, estas no son las únicas conclusiones que se pueden destacar del reporte de calificaciones por carrera, a continuación, se presentarán análisis específicos de cada variable utilizando métodos gráficos, por lo que, permitirán la presentación de nuevos enfoques y conclusiones dada la variable. Sin embargo, es necesario recordar al lector que las variables referidas a códigos del estudiante y el periodo no serán sujetas a ningún análisis por las razones descritas al comienzo de este capítulo.

Antes de comenzar el análisis, se presentará un código que ha servido para la mejor presentación y visualización de los gráficos de cada variable, el código es:

Ilustración 185

Función necesaria para la mejor visualización de los gráficos

```
def get_periodo(id_dataframe):  
    """  
    Devuelve el nombre del periodo dado el id del dataframe de la lista de dataframes de los reportes  
    """  
    nombre, data_id = id_dataframe.split("0")  
    if data_id == "1": return "Periodo 2021-01"  
    elif data_id == "2": return "Periodo 2021-02"  
    elif data_id == "3": return "Periodo 2022-01"  
    elif data_id == "4": return "Periodo 2022-02"
```

Ahora, para comenzar los análisis, se iniciará con la facultad a la que pertenece un determinado estudiante, para esta ocasión se utilizó la técnica de gráfico de barras debido a que existen varias categorías y sus proporciones en algunos puntos es desproporcionada por lo que el uso de un gráfico de pastel no permitiría un análisis coherente. El código y gráficos para el análisis de la variable son los siguientes:

Ilustración 186

Código que genera los gráficos para la variable facultad del reporte de Calificaciones por Carrera

```
# Ordenamos las categorías alfabéticamente
orden_categorias = sorted(set(calificaciones[0]["facultad"].unique()))

# Número de filas y columnas para los subplots
nrows = 2
ncols = 2

# Crear la figura y los ejes para los subplots
fig, axes = plt.subplots(nrows=nrows, ncols=ncols, figsize=(15, 10))

# Nombre de la columna de interés
columna_interes = 'facultad'

# Iterar a través de los dataframes y crear un gráfico para cada uno
for idx, (calificacion, ax) in enumerate(zip(calificaciones, axes.flat)):
    sns.countplot(x=columna_interes, data=calificacion, ax=ax, order=orden_categorias)
    ax.set_xlabel(f"{columna_interes}")
    ax.set_xticklabels(ax.get_xticklabels(), rotation=90)
    ax.set_ylabel('Cantidad')
    ax.set_title(f'Gráfico de barras de {columna_interes} ({get_periodo(calificacion.name)})')

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=1.2)

# Mostrar la figura
plt.show()
```

Ilustración 187

Gráfico de barras del primer y segundo periodo de la variable facultad del reporte de Calificaciones por Carrera

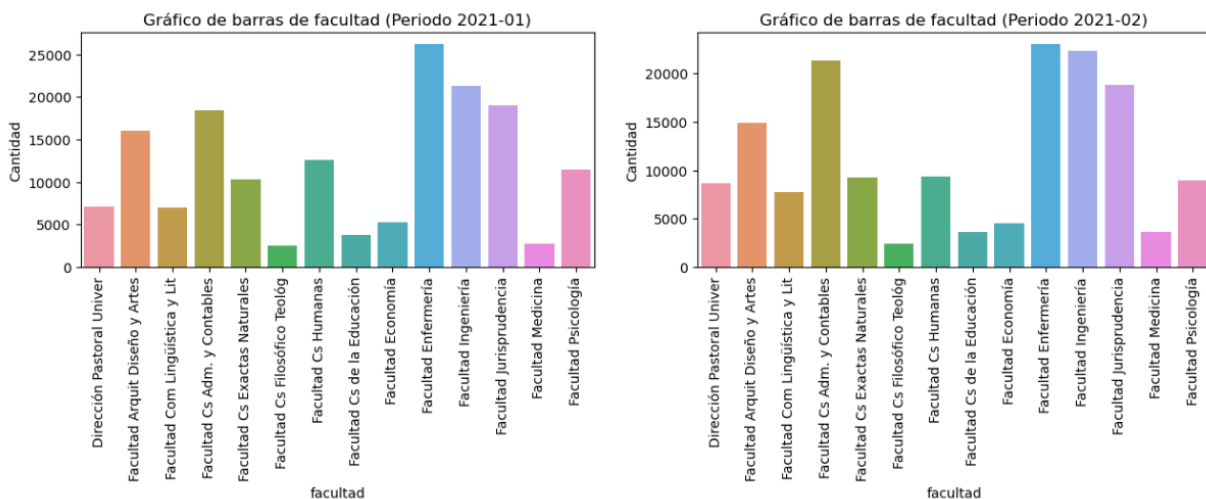
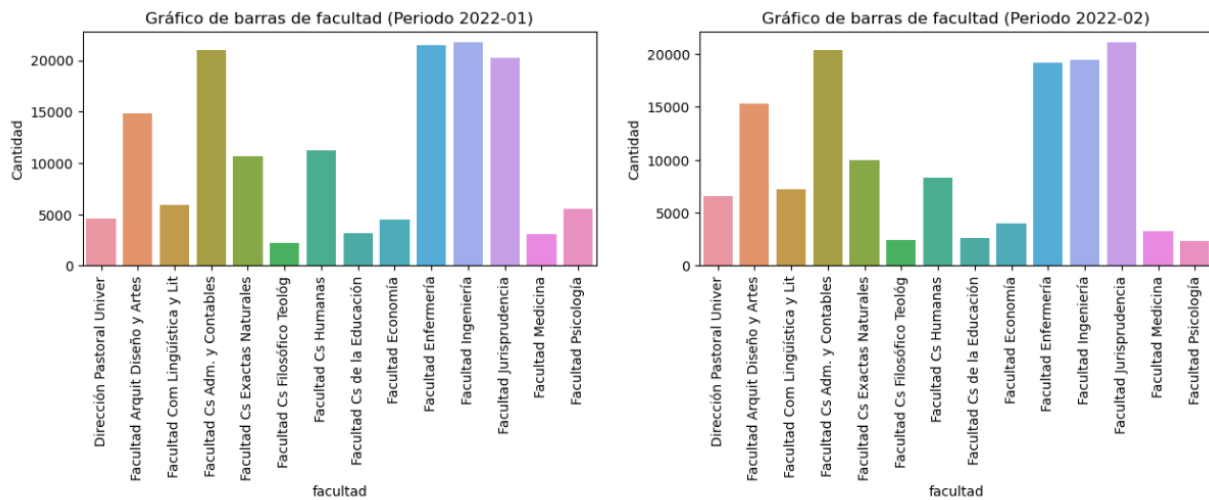


Ilustración 188

Gráfico de barras del tercer y cuarto periodo de la variable facultad del reporte de Calificaciones por Carrera



A partir de los gráficos de barras de la variable facultad, es posible destacar las siguientes conclusiones:

- Se puede observar que, a través de los periodos académicos, las facultades con mayor cantidad de estudiantes son:
 - Facultad de Enfermería
 - Facultad de Ingeniería
 - Facultad de Jurisprudencia
 - Facultad de Ciencias administrativas y contables
 - Facultad de Arquitectura Diseño y Artes
- Es observable que el patrón de distribución de estudiantes por facultad se mantiene igual por facultad a través de los periodos académicos, con leves aumentos y descensos.

Por otra parte, en segundo orden se visualizarán los gráficos de las carreras a las que pertenecen los distintos estudiantes, en esta ocasión, dado que existe una gran variedad de carreras universitarias en la Pontificia Universidad Católica del Ecuador y la proporción de estudiantes por carrera tiene una variación en proporción significativa, se utilizará la técnica de gráficos de barras. El código y los gráficos de esta variable podrán hallarse a continuación:

Ilustración 189

Código que genera los gráficos para la variable `descripcion_programa` del reporte de Calificaciones por Carrera

```
# Ordenamos las categorías alfabéticamente
orden_categorias = sorted(set(calificaciones[0]["descripcion_programa"].unique()))

# Número de filas y columnas para los subplots
nrows = 4
ncols = 1

# Crear la figura y los ejes para los subplots
fig, axes = plt.subplots(nrows=nrows, ncols=ncols, figsize=(15, 17))

# Nombre de la columna de interés
columna_interes = 'descripcion_programa'

# Iterar a través de los dataframes y crear un gráfico para cada uno
for idx, (calificacion, ax) in enumerate(zip(calificaciones, axes.flat)):
    sns.countplot(x=columna_interes, data=calificacion, ax=ax, order=orden_categorias)
    ax.set_xlabel(f"{columna_interes}")
    ax.set_xticklabels(ax.get_xticklabels(), rotation=90)
    ax.set_ylabel('Cantidad')
    ax.set_title(f'Gráfico de barras de {columna_interes} ({get_periodo(calificacion.name)})')

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=2.2, hspace=3.2)

# Mostrar la figura
plt.show()
```

Ilustración 190

Gráfico de barras del primer periodo de la variable descripcion_programa del reporte de Calificaciones por Carrera

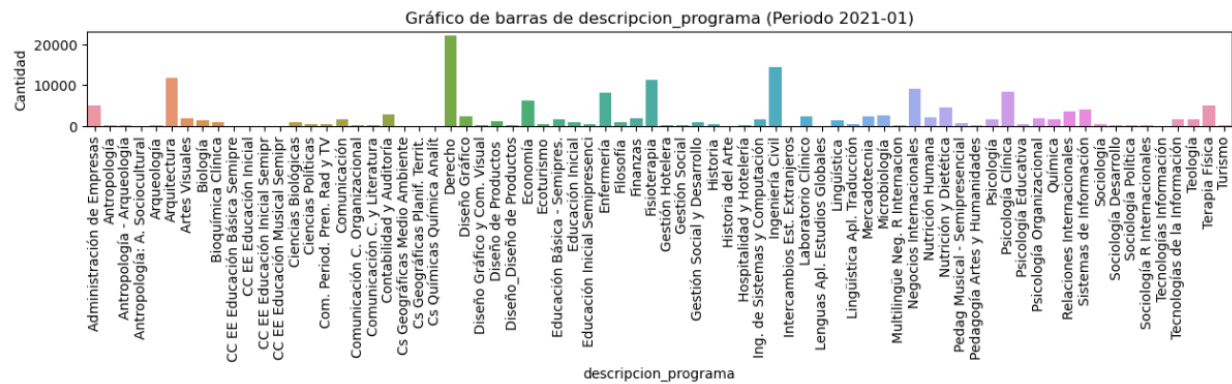


Ilustración 191

Gráfico de barras del segundo periodo de la variable descripcion_programa del reporte de Calificaciones por Carrera

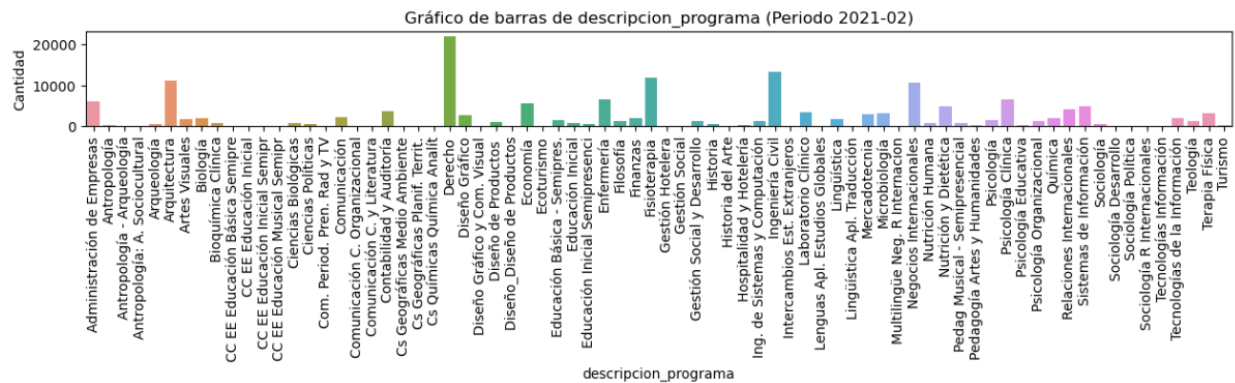


Ilustración 192

Gráfico de barras del tercer periodo de la variable descripcion_programa del reporte de Calificaciones por Carrera

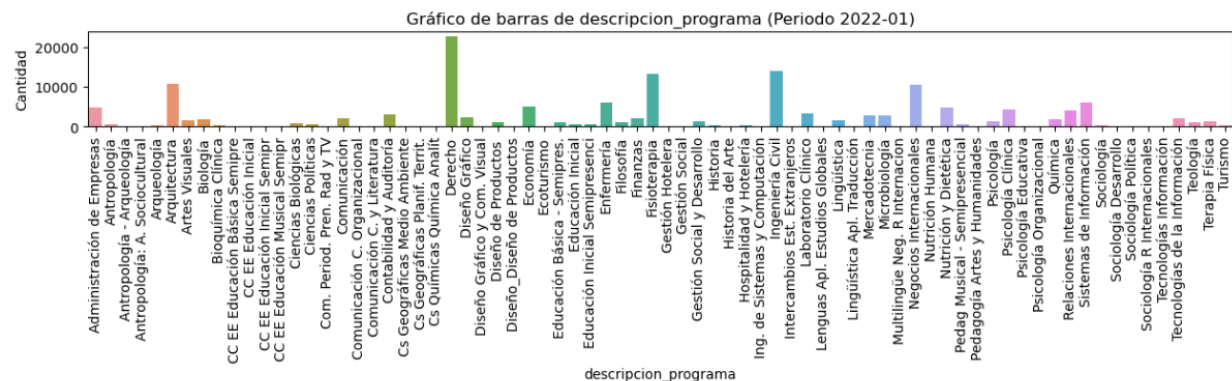
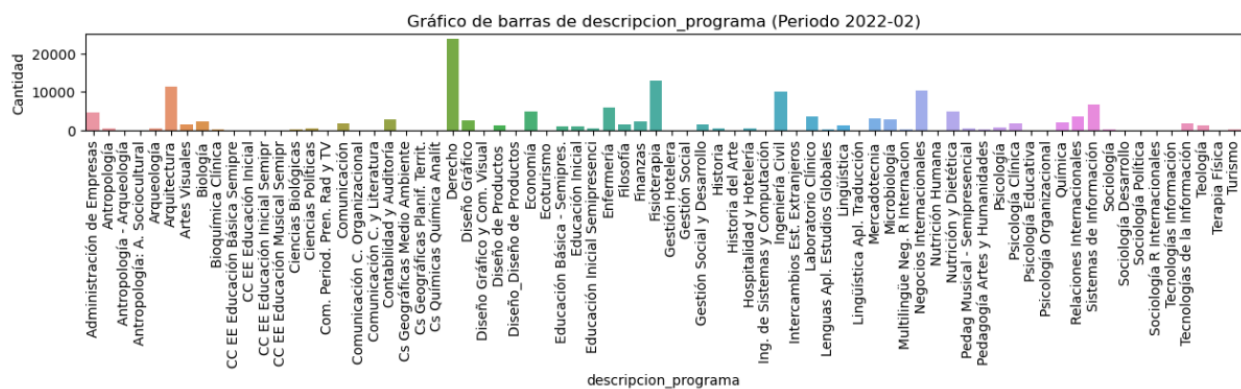


Ilustración 193

Gráfico de barras del cuarto periodo de la variable descripcion_programa del reporte de Calificaciones por Carrera



En síntesis, dado los gráficos presentados se puede concluir que como se mencionó previamente en las estadísticas descriptivas, la carrera con mayor cantidad de estudiantes es Derecho, seguida por Arquitectura, Ingeniería Civil, Fisioterapia e incluso Negocios Internacionales; esto indica que estas carreras tienen una mayor demanda entre los estudiantes universitarios. Adicionalmente, se puede observar un flujo similar de estudiantes en estas carreras a lo largo de los periodos considerados, por lo que sugiere una cierta estabilidad en la elección de carreras por parte de los estudiantes, independientemente del periodo académico.

En otro sentido, el tercer análisis que se realizará sobre las variables del reporte de calificaciones por carrera será la variable nota_final_SHATCKG, la cual, se debe recordar que se ha tratado como una variable categórica dado un análisis de sus valores únicos; sin embargo, dado que posee una considerable cantidad de datos únicos, se utilizará la técnica de gráfico de barras. A continuación, se podrá apreciar el código y los gráficos de la variable:

Ilustración 194

Código que genera los gráficos para la variable `nota_final_SHATCKG` del reporte de Calificaciones por Carrera

```
# Ordenamos las categorías alfabéticamente
orden_categorias = sorted(set(calificaciones[3]["nota_final_SHATCKG"].unique()))

# Número de filas y columnas para los subplots
nrows = 4
ncols = 1

# Crear la figura y los ejes para los subplots
fig, axes = plt.subplots(nrows=nrows, ncols=ncols, figsize=(15, 17))

# Nombre de la columna de interés
columna_interes = 'nota_final_SHATCKG'

# Iterar a través de los dataframes y crear un gráfico para cada uno
for idx, (calificacion, ax) in enumerate(zip(calificaciones, axes.flat)):
    sns.countplot(x=columna_interes, data=calificacion, ax=ax, order=orden_categorias)
    ax.set_xlabel(f"{columna_interes}")
    ax.set_xticklabels(ax.get_xticklabels(), rotation=90)
    ax.set_ylabel('Cantidad')
    ax.set_title(f'Gráfico de barras de {columna_interes} ({get_perodo(calificacion.name)})')

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=1.2, hspace=0.4)

# Mostrar la figura
plt.show()
```

Ilustración 195

Gráfico de barras del primer periodo de la variable `nota_final_SHATCKG` del reporte de Calificaciones por Carrera

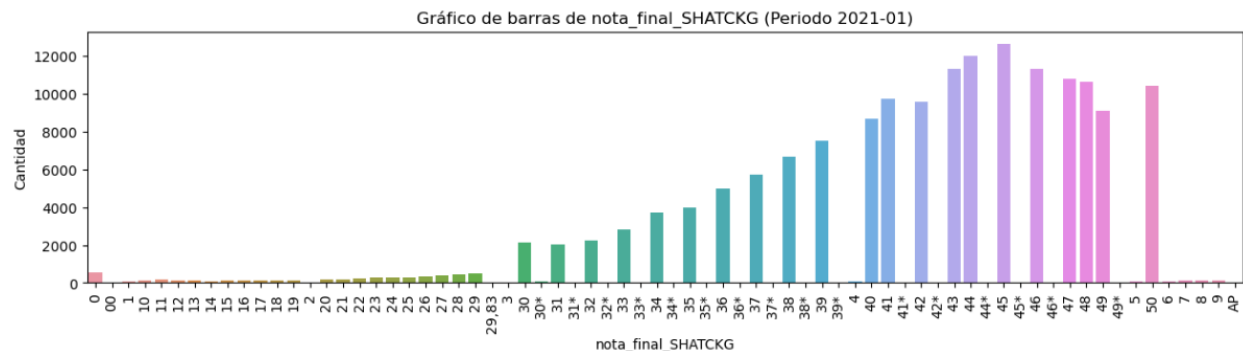


Ilustración 196

Gráfico de barras del segundo periodo de la variable nota_final_SHATCKG del reporte de Calificaciones por Carrera

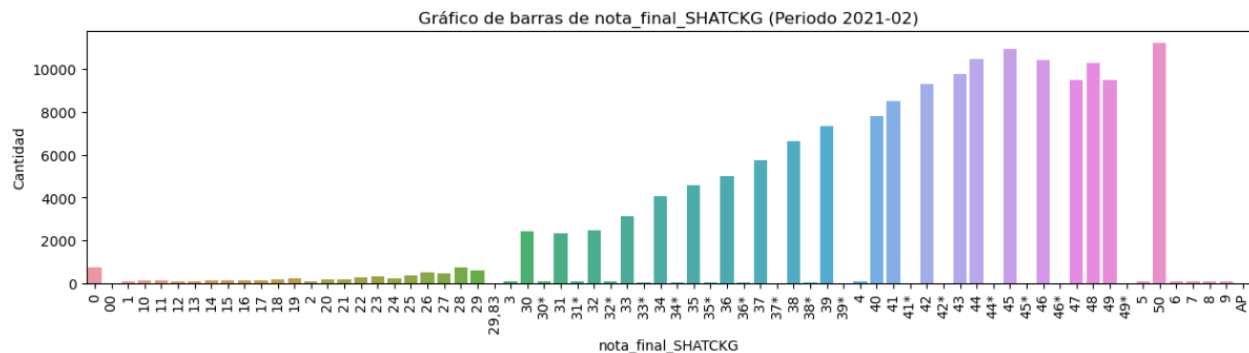


Ilustración 197

Gráfico de barras del tercer periodo de la variable nota_final_SHATCKG del reporte de Calificaciones por Carrera

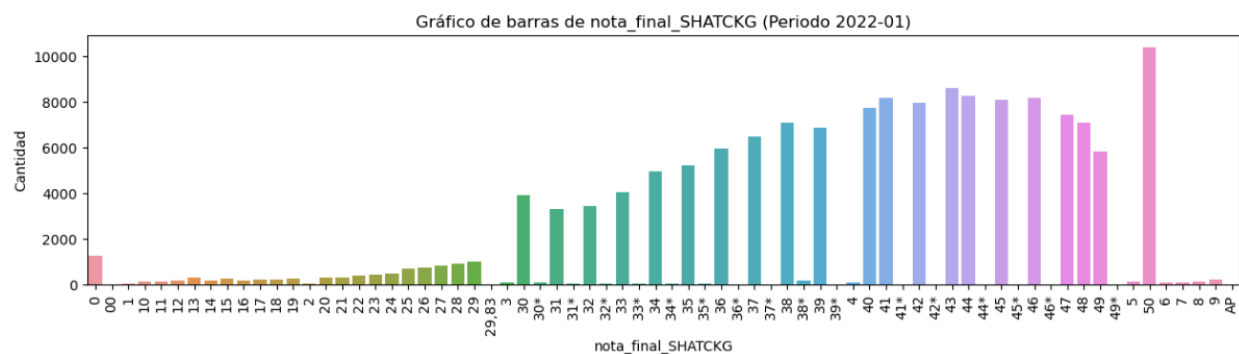
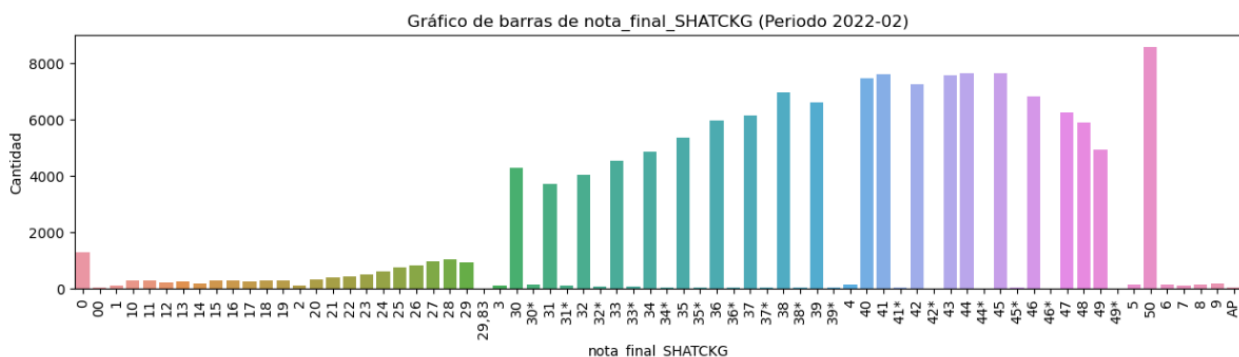


Ilustración 198

Gráfico de barras del cuarto periodo de la variable nota_final_SHATCKG del reporte de Calificaciones por Carrera



En resumen, la observación más relevante de estos gráficos es la presencia de valores como "AP" que se refieren a "Aprobado". Sin embargo, se observa que estos valores no siempre cumplen con la condición de aprobación, por lo que, debido a esta inconsistencia y considerando que ya se dispone de una variable que registra las calificaciones finales sin problemas significativos, lo más prudente sería descartar la utilidad de esta variable para evitar posibles confusiones o errores en el análisis.

Es decir, de manera concluyente de la exploración de las variables categóricas del reporte de calificaciones por carrera, se determinó que todas las variables son de interés y que requieren del análisis de relaciones para determinar finalmente su influencia sobre el rendimiento académico, sin embargo, la única excepción a este criterio fue la variable `nota_final_SHATCKG`, que, como se concluyó, esta variable no será considerada para futuros análisis.

En otro aspecto, para finalizar el capítulo de exploración de variables categóricas, es momento idóneo para realizar el análisis del reporte de becas, pero, teniendo en cuenta las consideraciones generales de que no se realizarán análisis más allá de las estadísticas descriptivas para las variables correspondientes al código de estudiantes y periodos.

Tomando en cuenta la premisa inicial, el análisis comenzará con el primero periodo del reporte de becas, cuyo código y observaciones serán observables a continuación:

Ilustración 199

Parte 1 de las Estadísticas Descriptivas del primer periodo del reporte de Calificaciones por Carrera

```
categoric_becas[0].describe()
```

	ID_BANNER	NACIONALIDAD	GENERO	ETNIA	TIPO_COLEGIO	PROGAMA_DESC	AREA_MATERIAS	NIVEL_POR_MATERIAS
count	9399	9255	9399	4099	2503	9399	9399	9350
unique	8607	27	3	8	5	74	13	12
top	E5195	Ecuatoriano	Mujer	Mestizo	Part R	Derecho	04	04
freq	4	9160	5070	3851	876	1159	1672	1672

Ilustración 200

Parte final de las Estadísticas Descriptivas del primer periodo del reporte de Calificaciones por Carrera

NIVEL_POR_ASIG_APROBADAS	CATEGORIA_BECA_SOCIOECONOMICA	TIENE_PAGO	TIENE_SALDO
7897	8568	9399	9399
10	5	2	2
8	C	SI	NO
1668	3708	9397	9349

A partir de las estadísticas descriptivas observadas, se destacaron las siguientes observaciones:

- Se puede observar que existen valores faltantes en el reporte debido a la cantidad de datos que posee cada variable, siendo las variables ETNIA y TIPO_COLEGIO las que encabezan este aspecto.
- Se puede observar que alrededor de todos los registros corresponden a estudiantes únicos, lo que demuestra fácilmente que el estudiante con mayor frecuencia tiene 4 apariciones.

- Con respecto a la nacionalidad, a pesar de que se tengan registros de 27 nacionalidades distintas, la gran mayoría de los estudiantes es ecuatoriano.
- Se puede observar que existe una mayoría de estudiante de género femenino, sin embargo, la diferencia no es un muy significativa con respecto al género masculino.
- Se puede observar que los estudiantes el periodo pertenecen a 5 diferentes tipos de colegio, pero la mayoría que posee beca socioeconómica ha asistido a un colegio particular religioso.
- La carrera más notoria es la de derecho de las 74 existentes, es decir, acumula un gran porcentaje en comparación a las demás.
- El registro correspondiente al área de materia en el que un estudiante se encuentra resulta irrelevante puesto que la asignatura que un estudiante curse no es un determinante de si va a poseer una beca socioeconómica o no, por lo que su contenido es engañoso.
- Se puede observar que la mayoría de los estudiantes del reporte pertenecen a cuarto semestre mientras que la mayoría de los registros muestran que han aprobado una asignatura de octavo semestre.
- Se puede observar que existen 5 tipos de beca socioeconómica y que la más común es la asignación del rango “C”, esto se debe principalmente a que en época de pandemia se asignó automáticamente la beca socioeconómica “C” a todos los estudiantes que ingresaron y solicitaron la misma.
- Se puede observar que únicamente 2 estudiante no pagaron su semestre, en contraste, se puede observar que 50 estudiantes poseen saldo pendiente.

En segundo lugar, dentro del contexto de la exploración de variables categóricas del reporte de becas, se visualizarán las estadísticas descriptivas del segundo periodo del reporte junto a sus observaciones:

Ilustración 201

Parte 1 de las Estadísticas Descriptivas del segundo periodo del reporte de Calificaciones por Carrera

```
categoric_becas[1].describe(include="object")
```

	ID_BANNER	NACIONALIDAD	GENERO	ETNIA	TIPO_COLEGIO	PROGAMA_DESC	AREA_MATERIAS	NIVEL_POR_MATERIAS
count	8564	8398	8564	2875	3180	8564	8564	8487
unique	8559	22	3	8	5	74	14	13
top	E3615	Ecuatoriano	Mujer	Mestizo	Part R	Derecho	05	05
freq	2	8316	4531	2701	1141	1088	1412	1412

Ilustración 202

Parte final de las Estadísticas Descriptivas del segundo periodo del reporte de Calificaciones por Carrera

NIVEL_POR_ASIG_APROBADAS	CATEGORIA_BECA_SOCIOECONOMICA	TIENE_PAGO	TIENE_SALDO
7156	0	8564	8564
10	0	2	2
7	NaN	SI	NO
1391	NaN	8559	8535

Tras el análisis de las estadísticas descriptivas visualizadas, se puede observar los siguientes comportamientos:

- Se puede observar que existen valores faltantes en el reporte debido a la cantidad de datos que posee cada variable, en esta ocasión, a pesar de que las variables ETNIA y TIPO_COLEGIO se mantienen con un importante número de valores faltantes, la

variable correspondiendo a los datos de las becas se encuentra en su totalidad vacía, esto, posiblemente se deba a que las asignaciones de becas se realizan una vez al año, es decir, cada dos periodos académicos.

- Se puede observar que alrededor de todos los registros corresponden a estudiantes únicos, lo que demuestra fácilmente que el estudiante con mayor frecuencia tiene 2 apariciones en esta ocasión.
- Con respecto a la nacionalidad, se presenta una disminución en la cantidad de nacionalidades, sin embargo, se mantiene que la gran mayoría de los estudiantes es ecuatoriana.
- Se puede observar que existe una mayoría de estudiante de género femenino, sin embargo, la diferencia sigue sin ser muy significativa con respecto al género masculino.
- Se puede observar que nuevamente los estudiantes el periodo pertenecen a 5 diferentes tipos de colegio, y, de manera similar al anterior periodo la mayoría que posee beca socioeconómica ha asistido a un colegio particular religioso.
- La carrera más notoria nuevamente es la de derecho de las 74 existentes, es decir, acumula un gran porcentaje en comparación a las demás.
- Con respecto a las conclusiones de la variable que AREA_MATERIAS, se mantiene la misma conclusión en la que sus datos son engañosos y no poseen sentido para su contexto.

- Se puede observar que la mayoría de los estudiantes del reporte pertenecen a quinto semestre mientras que la mayoría de los registros muestran que han aprobado una asignatura de séptimo semestre, es evidente que la mayor cantidad de estudiante con beca socioeconómica tienen asignaturas de niveles altos aprobadas.
- Dado que no existen registros de las becas socioeconómicas del reporte, es probable que los registros correspondientes a esta variable no presenten utilidad si evalúa su influencia en el modelo de machine learning planteado.
- Se puede observar que 5 estudiante no pagaron su semestre y, en contraste, se puede observar que 29 estudiantes poseen un saldo pendiente.

Como tercer término, es momento de presentar las estadísticas descriptivas del tercer periodo del reporte:

Ilustración 203

Parte 1 de las Estadísticas Descriptivas del tercer periodo del reporte de Calificaciones por Carrera

```
categoric_becas[2].describe(include="object")
```

	ID_BANNER	NACIONALIDAD	GENERO	ETNIA	TIPO_COLEGIO	PROGAMA_DESC	AREA_MATERIAS	NIVEL_POR_MATERIAS
count	8336	8152	8335	2309	3501	8335	8336	8223
unique	8232	23	3	8	5	72	14	13
top	E2113	Ecuatoriano	Mujer	Mestizo	Part R	Derecho	06	06
freq	4	8071	4366	2166	1257	1111	1437	1437

Ilustración 204

Parte final de las Estadísticas Descriptivas del tercer periodo del reporte de Calificaciones por Carrera

NIVEL_POR_ASIG_APROBADAS	CATEGORIA_BECA_SOCIOECONOMICA	TIENE_PAGO	TIENE_SALDO
6971	8187	8334	8334
10	5	2	2
7	C	SI	NO
1367	4287	8328	8333

A partir de las estadísticas descriptivas exhibidas, se puede observar lo siguiente:

- Se puede observar que existen valores faltantes en el reporte debido a la cantidad de datos que posee cada variable, en esta ocasión, a pesar de que las variables ETNIA y TIPO_COLEGIO se mantienen con un importante número de valores faltantes, la variable correspondiendo a los datos de las becas ya no se encuentra con una fuerte cantidad de valores faltantes, en su lugar, ha aumentado el de la variable NIVEL_POR_ASIG_APROBADAS.
- Se puede observar que alrededor de todos los registros corresponden a estudiantes únicos, lo que demuestra fácilmente que el estudiante con mayor frecuencia tiene 4 apariciones nuevamente.
- Con respecto a la nacionalidad, se presenta nuevamente una disminución en la cantidad de nacionalidades, sin embargo, se mantiene que la gran mayoría de los estudiantes es ecuatoriana.
- Se puede observar que existe una mayoría de estudiante de género femenino, sin embargo, la diferencia sigue sin ser muy significativa con respecto al género

masculino. Es decir, a lo largo de los periodos se sigue manteniendo una paridad en cuestiones de genero con respecto a la asignación de becas socioeconómicas.

- Se puede observar que nuevamente los estudiantes el periodo pertenecen a 5 diferentes tipos de colegio, y, de manera similar al anterior periodo la mayoría que posee beca socioeconómica ha asistido a un colegio particular religioso.
- La carrera más notoria nuevamente es la de derecho de las 72 existentes en este periodo, es decir, acumula un gran porcentaje en comparación a las demás.
- Con respecto a las conclusiones de la variable que AREA_MATERIAS, se mantiene la misma conclusión en la que sus datos son engañosos y no poseen sentido para su contexto.
- Se puede observar que la mayoría de los estudiantes del reporte pertenecen a sexto semestre mientras que la mayoría de los registros muestran que han aprobado una asignatura de sexto semestre, esto demuestra claramente como cada periodo, la mayoría de estudiantes que poseía beca en periodos anteriores la ha mantenido y, por lo tanto, su semestre actual se eleva, en otras palabras, el nivel académico más frecuente actualmente corresponde a los mismos estudiantes del nivel académico más frecuente de los periodos anteriores.
- Dado que en esta ocasión volvieron a aparecer los registros de las becas, se puede corroborar la trabajo de titulación de que la asignación de becas socioeconómicas se realiza una vez cada dos periodos.
- Se puede observar que 6 estudiante no pagaron su semestre y, en contraste, se puede observar que exclusivamente 1 estudiante posee un saldo pendiente.

Finalmente, las estadísticas descriptivas del cuarto y último periodo del reporte de becas son las siguientes:

Ilustración 205

Parte 1 de las Estadísticas Descriptivas del cuarto periodo del reporte de Calificaciones por Carrera

```
categoric_becas[3].describe(include="object")
```

	ID_BANNER	NACIONALIDAD	GENERO	ETNIA	TIPO_COLEGIO	PROGAMA_DESC	AREA_MATERIAS	NIVEL_POR_MATERIAS
count	8255	8018	8255	1437	4332	8255	8255	8062
unique	8253	25	3	8	5	71	13	12
top	E218	Ecuatoriano	Mujer	Mestizo	Part R	Derecho	07	07
freq	2	7939	4312	1339	1623	1188	1405	1405

Ilustración 206

Parte final de las Estadísticas Descriptivas del cuarto periodo del reporte de Calificaciones por Carrera

NIVEL_POR_ASIG_APROBADAS	CATEGORIA_BECA_SOCIOECONOMICA	TIENE_PAGO	TIENE_SALDO
6878	0	8255	8255
10	0	2	2
7	NaN	SI	NO
1079	NaN	8253	8253

A partir de las últimas estadísticas descriptivas se pueden destacar los siguientes aspectos:

- Se puede observar que existen valores faltantes en el reporte debido a la cantidad de datos que posee cada variable, en esta ocasión, a pesar de que las variables ETNIA y TIPO_COLEGIO se mantienen con un importante número de valores faltantes, la variable correspondiendo a los datos de las becas nuevamente se encuentra en su

totalidad vacía, esto por las razones descubiertas y tratadas en las observaciones de los anteriores periodos.

- Se puede observar que alrededor de todos los registros corresponden a estudiantes únicos, lo que demuestra fácilmente que el estudiante con mayor frecuencia tiene 2 apariciones de igual forma que para el segundo periodo.
- Con respecto a la nacionalidad, se presenta un pequeño en la cantidad de nacionalidades, sin embargo, se mantiene que la gran mayoría de los estudiantes es ecuatoriana.
- Se puede observar que existe una mayoría de estudiante de género femenino, sin embargo, la diferencia sigue sin ser muy significativa con respecto al género masculino. En general, en todos los periodos se mostró un comportamiento parejo en este sentido con respecto a la asignación de becas.
- Se puede observar que nuevamente los estudiantes el periodo pertenecen a 5 diferentes tipos de colegio, y, de manera similar al anterior periodo la mayoría que posee beca socioeconómica ha asistido a un colegio particular religioso.
- La carrera más notoria nuevamente es la de derecho de las 71 existentes, es decir, acumula un gran porcentaje en comparación a las demás.
- Con respecto a las conclusiones de la variable que AREA_MATERIAS, se mantiene la misma conclusión en la que sus datos son engañosos y no poseen sentido para su contexto.
- Se puede observar que, en sentido del nivel académico de los estudiantes con beca, cumple poseer el mismo sentido que en el anterior periodo.

- Dado que nuevamente no existen registros de las becas socioeconómicas del reporte, es probable que los registros correspondientes a esta variable no presenten utilidad si evalúa su influencia en el modelo de machine learning planteado.
- Se puede observar que 2 estudiante no pagaron su semestre y, en contraste, se puede observar que 2 estudiantes poseen un saldo pendiente.

De manera concluyente, a partir de la finalización de los análisis de las estadísticas descriptivas, se puede concluir lo siguiente del reporte de becas:

- Datos Faltantes: Los campos de ETNIA y TIPO_COLEGIO tienden a tener muchos datos faltantes en todos los periodos, por otra parte, el campo de los datos de las becas presenta un patrón de alternancia, siendo vacío en cada segundo periodo debido a que las asignaciones de becas se realizan cada dos periodos académicos. Esto indica la necesidad de mejorar la recopilación de datos para estos campos.
- Nacionalidad: La mayoría de los estudiantes son ecuatorianos, independientemente del periodo.
- Género: Existe una ligera mayoría de estudiantes femeninas, aunque la diferencia con el número de estudiantes masculinos no es significativa, lo que sugiere que la asignación de becas es relativamente equitativa en términos de género.
- Tipo de Colegio: La mayoría de los estudiantes que reciben becas han asistido a colegios particulares religiosos en todos los periodos, lo que parece curioso considerando que la universidad es de igual manera religiosa, lo que sugiere que es probable que exista una cierta relación.
- Carrera: La carrera de derecho parece ser la más popular entre los estudiantes que reciben becas en todos los periodos, por lo que se puede entender que los estudiantes

de derecho tienen una alta demanda de apoyo financiero o que la beca está dirigida hacia esa carrera en particular.

- **Semestre y Materias Aprobadas:** Hay una tendencia hacia estudiantes de semestres superiores (4to, 5to, 6to) que han aprobado asignaturas de niveles más altos, lo que indica que los estudiantes que reciben becas suelen mantenerlas a lo largo de su carrera universitaria.
- **Asignación de Becas:** La beca socioeconómica "C" es la más común en todos los registros, esto se debe principalmente a que esta fue la única asignada durante el periodo de pandemia.
- **Pago del Semestre:** A lo largo de los periodos, hay una pequeña cantidad de estudiantes que no pagaron su semestre, mientras que una mayor cantidad tiene un saldo pendiente.

Ahora, una vez realizado el análisis de las estadísticas descriptivas del reporte de becas y haber recolectado aspectos interesantes y curiosos de los datos de manera preliminar, es momento de realizar un análisis más profundo de las variables, sin embargo, es prudente recordar nuevamente al lector que se obviara el análisis de las variables que presenten códigos de estudiantes y periodos, no obstante, de manera adicional a esta premisa, en esta ocasión no se realizará el análisis de la variable `AREA_MATERIAS` por causa de las conclusiones de la análisis de las estadísticas descriptivas realizadas.

Antes de comenzar, se realizó una función que permitió una muestra decente y coherente de los gráficos para el análisis, este código fue el siguiente:

Ilustración 207

Función necesaria para la visualización coherente de los gráficos de los reportes

```
def get_periodo(id_dataframe):  
    '''  
    Devuelve el nombre del periodo dado el id del dataframe de la lista de dataframes de los reportes  
    '''  
    nombre, data_id = id_dataframe.split("0")  
    if data_id == "1": return "Periodo 2021-01"  
    elif data_id == "2": return "Periodo 2021-02"  
    elif data_id == "3": return "Periodo 2022-01"  
    elif data_id == "4": return "Periodo 2022-02"
```

Una vez mencionada la función que permitió un análisis más ordenado de las variables, el análisis comenzará enfocado en la nacionalidad de los estudiantes, para esta ocasión se utilizarán gráficos de barras debido a la inmensa desproporción de los datos de una variable con otra. El código y sus conclusiones se pueden encontrar a continuación:

Ilustración 208

Código que genera los gráficos para la variable NACIONALIDAD del reporte de Becas

```
# Número de filas y columnas para los subplots  
nrows = 2  
ncols = 2  
  
# Crear la figura y los ejes para los subplots  
fig, axes = plt.subplots(nrows=nrows, ncols=ncols, figsize=(15, 10))  
  
# Nombre de la columna de interés  
columna_interes = 'NACIONALIDAD'  
  
# Iterar a través de los dataframes y crear un gráfico para cada uno  
for idx, (beca, ax) in enumerate(zip(becas, axes.flat)):  
    sns.countplot(x=columna_interes, data=beca, ax=ax)  
    ax.set_xlabel(f"{columna_interes}")  
    ax.set_xticklabels(ax.get_xticklabels(), rotation=90)  
    ax.set_ylabel('Cantidad')  
    ax.set_title(f'Gráfico de barras de {columna_interes} ({get_periodo(beca.name)})')  
  
# Ajustar el espacio entre los subplots  
plt.subplots_adjust(wspace=0.2, hspace=1.2)  
  
# Mostrar la figura  
plt.show()
```

Ilustración 209

Gráfico de barras del primer y segundo periodo de la variable NACIONALIDAD del reporte de Becas

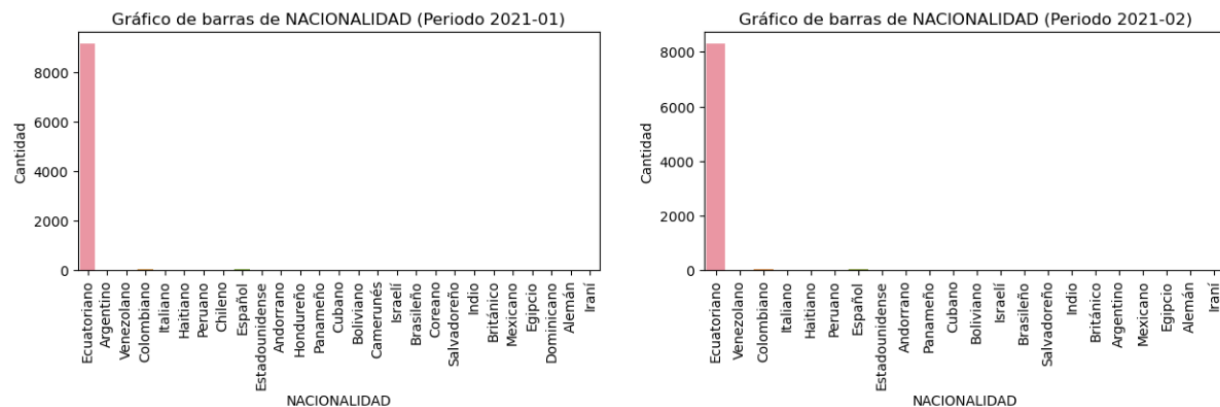
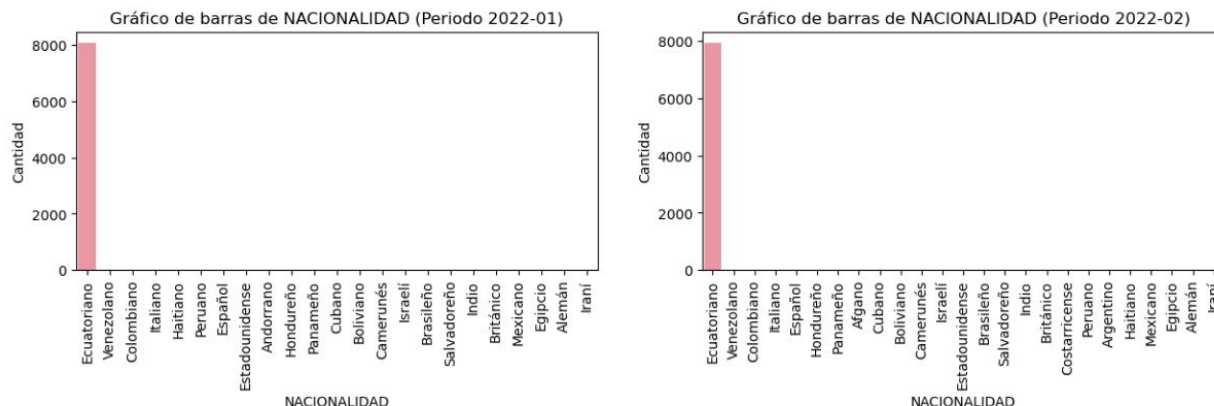


Ilustración 210

Gráfico de barras del tercer y cuarto periodo de la variable NACIONALIDAD del reporte de Becas



A partir de una visualización analítica de los gráficos se puede concretar que la mayor distribución de los estudiantes es de nacionalidad ecuatoriana en todos los periodos, adicionalmente, se demuestra que a pesar de que de las demás nacionalidades igualmente tengan registros, estos no representan una cantidad visible en los gráficos por lo que realmente la variable no resulta de utilidad para un modelo de machine learning, esto es debido al gran sesgo que existe en la distribución de estos registros.

Por otro lado, es momento de realizar el análisis del género de los estudiantes, para esta ocasión es factible el uso de la técnica del grafico de pastel debido a que permitirá un análisis más certero y brindará más información. El código, los gráficos y sus conclusiones se podrán encontrar a continuación:

Ilustración 211

Código que genera los gráficos para la variable GENERO del reporte de Becas

```
# Se especifica el número de filas y columnas para los subplots
nrows = 2
ncols = 2

# Se crea la figura y los ejes de los subplots con el tamaño especificado
fig, axes = plt.subplots(nrows=nrows, ncols=ncols, figsize=(15, 10))

# Se establece el nombre de la columna de interés
columna_interes = 'GENERO'

# Se itera a través de los dataframes y se crea un gráfico de pie para cada uno
for idx, (beca, ax) in enumerate(zip(becas, axes.flat)):
    # Se obtiene la cantidad de cada categoría en la columna de interés
    counts = beca[columna_interes].value_counts()
    # Se obtienen las etiquetas y los tamaños para el gráfico de pie
    labels = counts.index
    sizes = counts.values
    # Se crea el gráfico de pie en los ejes actuales
    ax.pie(sizes, labels=labels, autopct='%1.1f%%', startangle=90)
    # Se establece el título para cada subplot
    ax.set_title(f'Gráfico de pastel de {columna_interes} ({get_periodo(beca.name)})')

# Se ajusta el espacio entre los subplots para que no se superpongan
plt.subplots_adjust(wspace=0.2, hspace=0.4)

# Se muestra la figura con los subplots
plt.show()
```

Ilustración 212

Gráfico de pastel del primer y segundo periodo de la variable GENERO del reporte de Becas

Gráfico de pastel de GENERO (Periodo 2021-01)

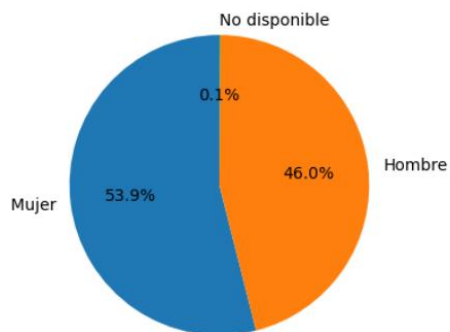


Gráfico de pastel de GENERO (Periodo 2021-02)

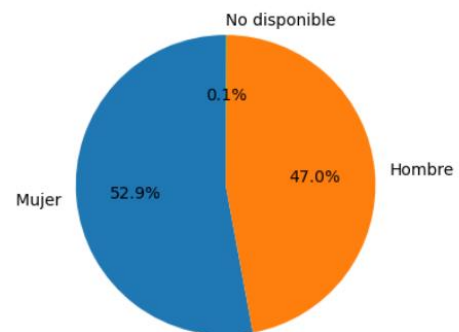


Ilustración 213

Gráfico de pastel del tercer y cuarto periodo de la variable GENERO del reporte de Becas

Gráfico de pastel de GENERO (Periodo 2022-01)

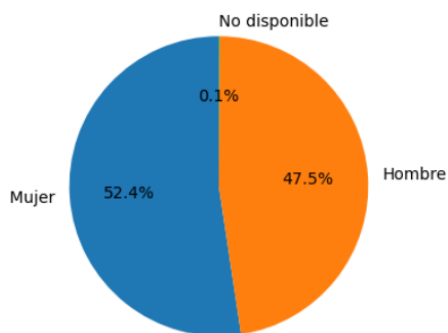
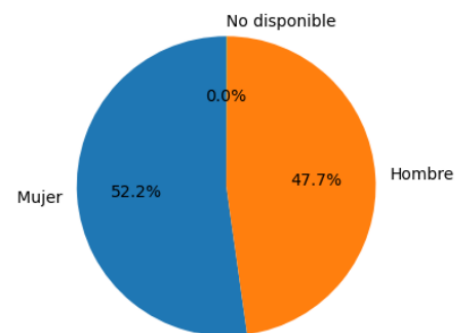


Gráfico de pastel de GENERO (Periodo 2022-02)



A partir de la visualización de los intuitivos gráficos de pastel, se puede concluir que en todos los periodos existe una mayor cantidad, pero no tan significativa, de mujeres que de hombres en el reporte de becas; mientras que, existe un mínimo porcentaje de estudiantes que no han reportado su género.

Desde otra perspectiva, el análisis de la etnia se realizará con la técnica de gráficos de barras debido a la desproporción de los datos. Su código, gráficos y conclusiones se podrá encontrar a continuación:

Ilustración 214

Código que genera los gráficos para la variable ETNIA del reporte de Becas

```
# Número de filas y columnas para los subplots
nrows = 2
ncols = 2

# Crear la figura y los ejes para los subplots
fig, axes = plt.subplots(nrows=nrows, ncols=ncols, figsize=(15, 10))

# Nombre de la columna de interés
columna_interes = 'ETNIA'

# Iterar a través de los dataframes y crear un gráfico para cada uno
for idx, (beca, ax) in enumerate(zip(becas, axes.flat)):
    sns.countplot(x=columna_interes, data=beca, ax=ax)
    ax.set_xlabel(f"{columna_interes}")
    ax.set_xticklabels(ax.get_xticklabels(), rotation=90)
    ax.set_ylabel('Cantidad')
    ax.set_title(f'Gráfico de barras de {columna_interes} ({get_periodo(beca.name)})')

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.6)

# Mostrar la figura
plt.show()
```

Ilustración 215

Gráfico de barras del primer y segundo periodo de la variable ETNIA del reporte de Becas

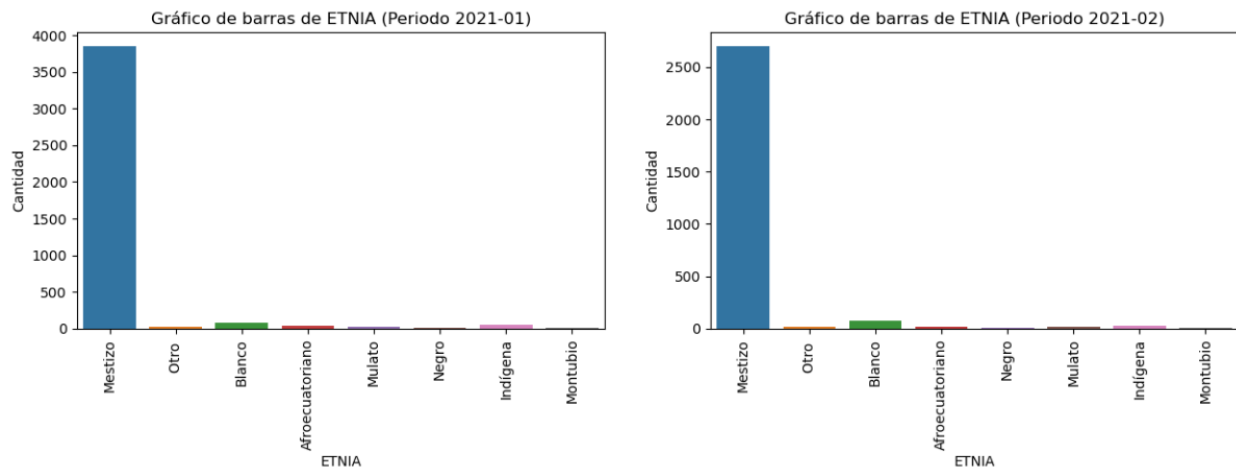
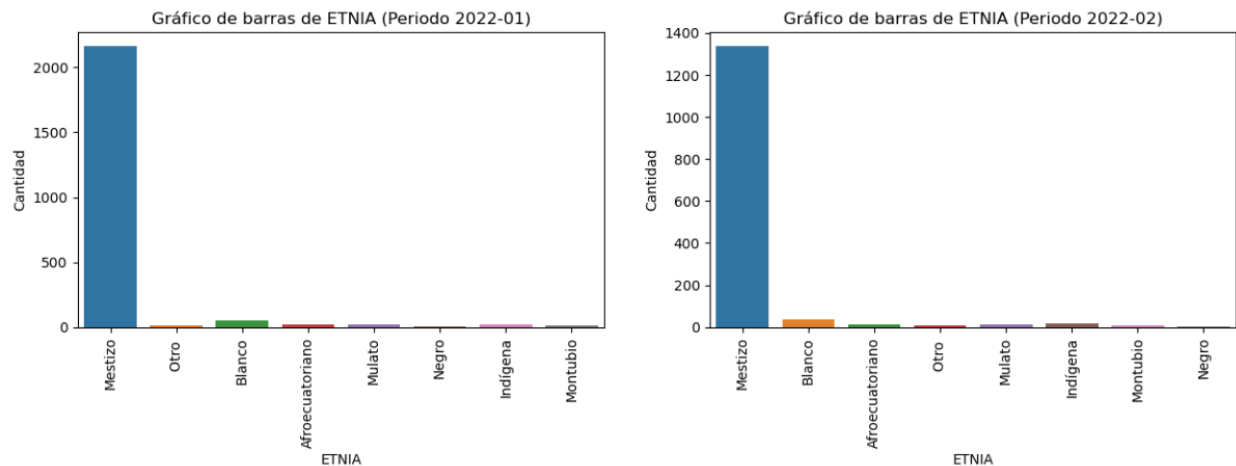


Ilustración 216

Gráfico de barras del tercer y cuarto periodo de la variable ETNIA del reporte de Becas



Como conclusión, a diferencia de la nacionalidad, se observa claramente que las diferentes etnias tienen un impacto significativo, aunque mínimo, no insignificante; es decir, a pesar de que existe un marcado sesgo en los datos, es posible que exista una relación entre el rendimiento académico y la etnia. Por lo tanto, esta variable podría considerarse para ser utilizada en un modelo con el fin de explorar y analizar su influencia en el rendimiento académico de los estudiantes.

Desde otro punto de vista, con respecto al análisis del tipo de colegio al que asistieron los estudiantes, se empleará la técnica del gráfico de pastel porque permitirá un destacar conclusiones con mayor efectividad. El código, gráficos y conclusiones de la variable TIPO_COLEGIO se podrá encontrar a continuación:

Ilustración 217

Código que genera los gráficos para la variable TIPO_COLEGIO del reporte de Becas

```
# Número de filas y columnas para los subplots
nrows = 2
ncols = 2

# Crear la figura y los ejes para los subplots
fig, axes = plt.subplots(nrows=nrows, ncols=ncols, figsize=(15, 10))

# Nombre de la columna de interés
columna_interes = 'TIPO_COLEGIO'

# Iterar a través de los dataframes y crear un gráfico para cada uno
for idx, (beca, ax) in enumerate(zip(becas, axes.flat)):
    # Obtener la cantidad de cada categoría en la columna de interés
    counts = beca[columna_interes].value_counts()
    # Obtener las etiquetas y tamaños para el gráfico de pie
    labels = counts.index
    sizes = counts.values
    # Crear el gráfico de pie en los ejes actuales
    ax.pie(sizes, labels=labels, autopct='%1.1f%%', startangle=90)
    # Establecer el título para cada subplot
    ax.set_title(f'Gráfico de pastel de {columna_interes} ({get_periodo(beca.name)})')

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.4)

# Mostrar la figura
plt.show()
```

Ilustración 218

Gráfico de barras del primer y segundo periodo de la variable TIPO_COLEGIO del reporte de Becas

Gráfico de pastel de TIPO_COLEGIO (Periodo 2021-01)

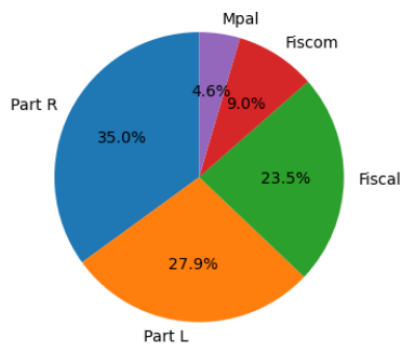


Gráfico de pastel de TIPO_COLEGIO (Periodo 2021-02)

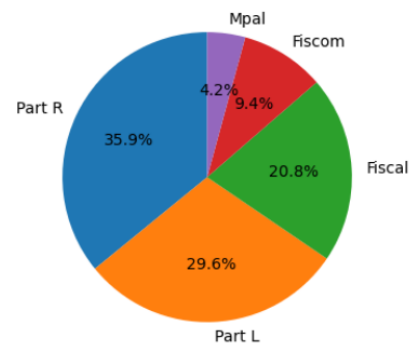


Ilustración 219

Gráfico de barras del tercer y cuarto periodo de la variable TIPO_COLEGIO del reporte de Becas

Gráfico de pastel de TIPO_COLEGIO (Periodo 2022-01)

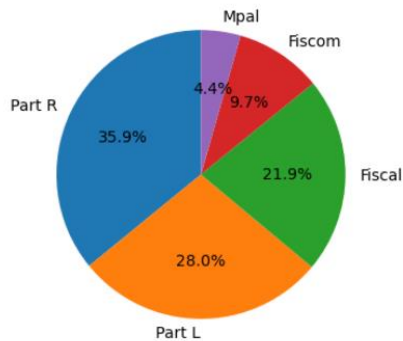
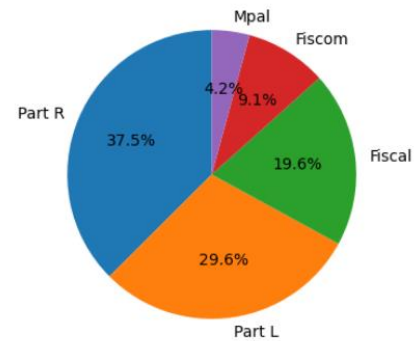


Gráfico de pastel de TIPO_COLEGIO (Periodo 2022-02)



A partir de los gráficos de pastel visualizados, se puede concluir que el comportamiento de los registros no varía dependiendo del periodo, la cantidad de estudiantes de cada tipo de colegio mantienen un mismo patrón de comportamiento a través del tiempo.

Por otra parte, el análisis correspondiente a las carreras universitarias de los estudiantes se realizará con la técnica de gráfico de barras debido a la gran cantidad de carreras existentes en la universidad. El código, gráficos y conclusión se podrá apreciar a continuación:

Ilustración 220

Código que genera los gráficos para la variable PROGAMA_DESC del reporte de Becas

```
# Ordenamos las categorías alfabéticamente
orden_categorias = sorted(set(becas[0]["PROGAMA_DESC"].unique()))

# Número de filas y columnas para los subplots
nrows = 4
ncols = 1

# Crear la figura y los ejes para los subplots
fig, axes = plt.subplots(nrows=nrows, ncols=ncols, figsize=(15, 17))

# Nombre de la columna de interés
columna_interes = 'PROGAMA_DESC'

# Iterar a través de los dataframes y crear un gráfico para cada uno
for idx, (beca, ax) in enumerate(zip(becas, axes.flat)):
    sns.countplot(x=columna_interes, data=beca, ax=ax, order=orden_categorias)
    ax.set_xlabel(f"{columna_interes}")
    ax.set_xticklabels(ax.get_xticklabels(), rotation=90)
    ax.set_ylabel('Cantidad')
    ax.set_title(f'Gráfico de barras de {columna_interes} ({get_periodo(beca.name)})')

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=2.2, hspace=3.2)

# Mostrar la figura
plt.show()
```

Ilustración 221

Gráfico de barras del primer periodo de la variable PROGAMA_DESC del reporte de Becas

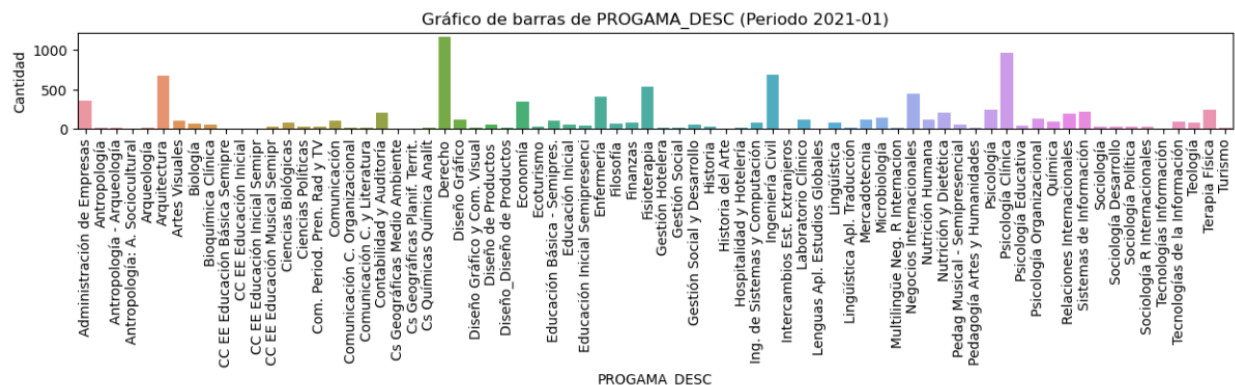


Ilustración 222

Gráfico de barras del segundo periodo de la variable PROGAMA_DESC del reporte de Becas

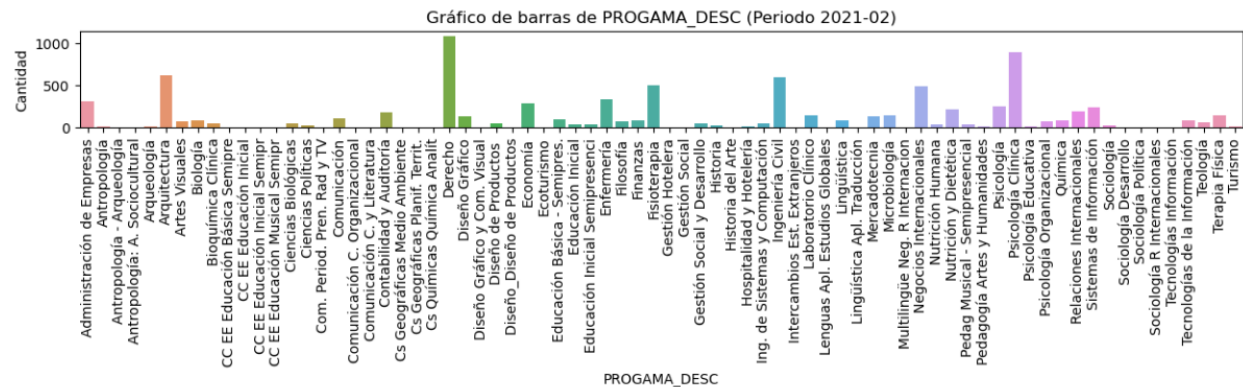


Ilustración 223

Gráfico de barras del tercer periodo de la variable PROGAMA_DESC del reporte de Becas

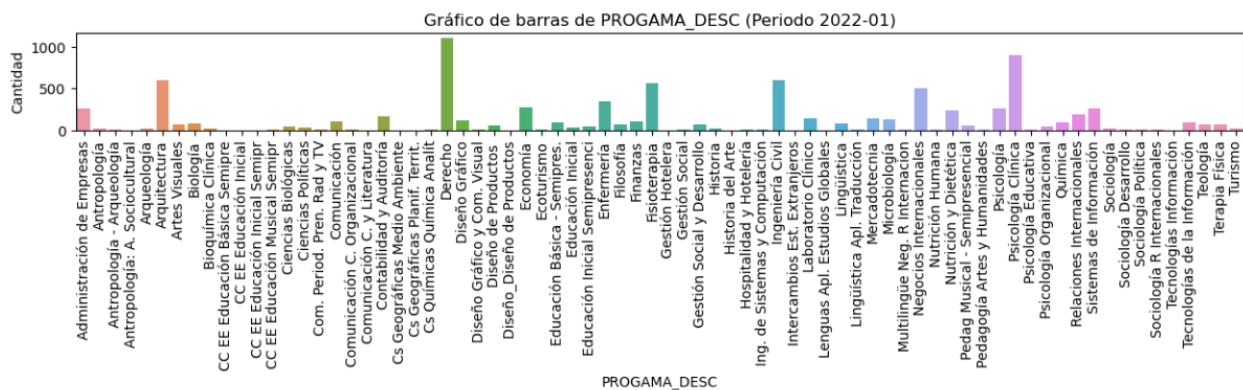
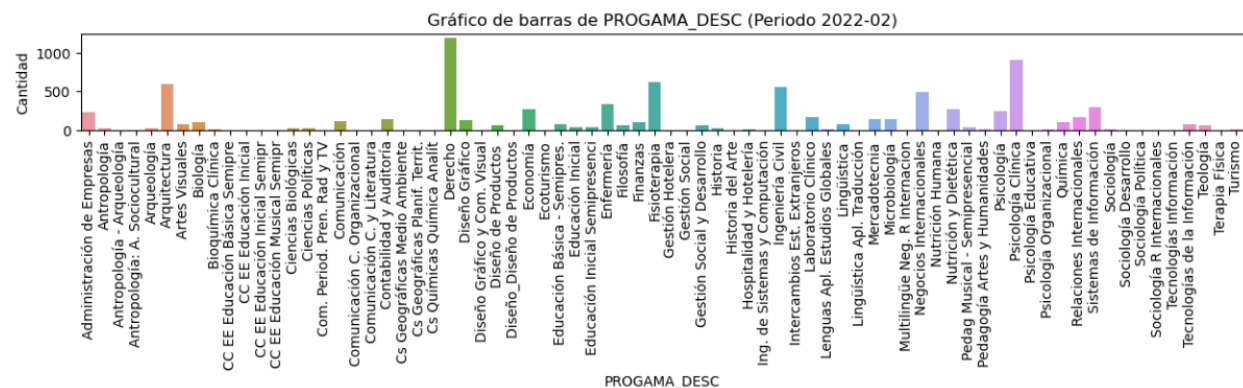


Ilustración 224

Gráfico de barras del cuarto periodo de la variable PROGAMA_DESC del reporte de Becas



A partir de los gráficos visualizados de las carreras universitarias de los estudiantes en el reporte de becas, se puede concluir que las carreras que poseen una mayor cantidad de estudiantes con becas son las de Derecho, Psicología Clínica y Arquitectura; adicionalmente, se puede observar, de igual modo que anteriores variables, un comportamiento similar sin importar a través del tiempo, es decir, de los periodos académicos.

Ahora, ha llegado el momento de analizar el nivel académico de los estudiantes con beca, para ello se emplearán gráficos de barras debido a que existen una moderada cantidad de niveles. El código, gráfico y conclusiones se podrán hallar a continuación:

Ilustración 225

Código que genera los gráficos para la variable NIVEL_POR_MATERIAS del reporte de Becas

```
# Número de filas y columnas para los subplots
nrows = 2
ncols = 2

# Crear la figura y los ejes para los subplots
fig, axes = plt.subplots(nrows=nrows, ncols=ncols, figsize=(15, 10))

# Nombre de la columna de interés
columna_interes = 'NIVEL_POR_MATERIAS'

# Iterar a través de los dataframes y crear un gráfico para cada uno
for idx, (beca, ax) in enumerate(zip(becas, axes.flat)):
    sns.countplot(x=columna_interes, data=beca, ax=ax)
    ax.set_xlabel(f"{columna_interes}")
    ax.set_ylabel('Cantidad')
    ax.set_title(f'Gráfico de barras de {columna_interes} ({get_perodo(beca.name)})')

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.4)

# Mostrar la figura
plt.show()
```

Ilustración 226

Gráfico de barras del primer y segundo periodo de la variable NIVEL_POR_MATERIAS del reporte de Becas

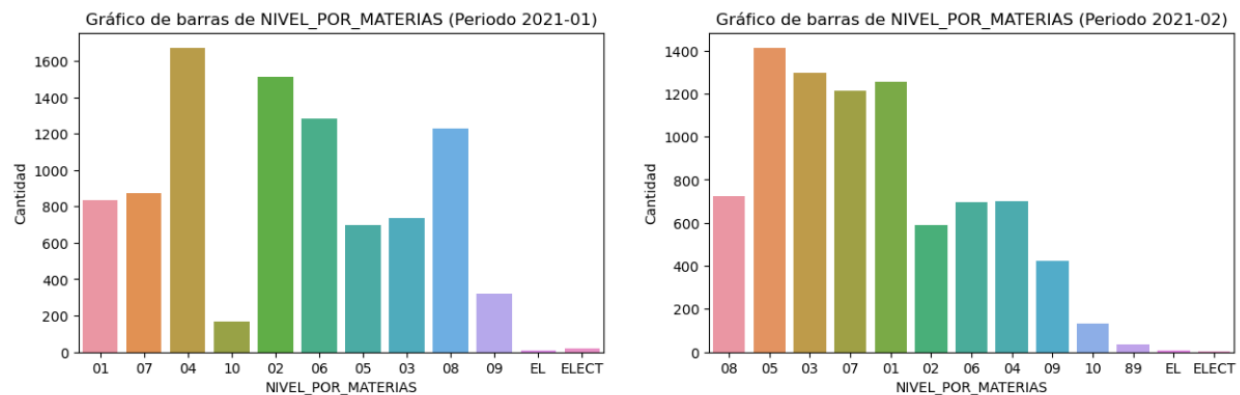
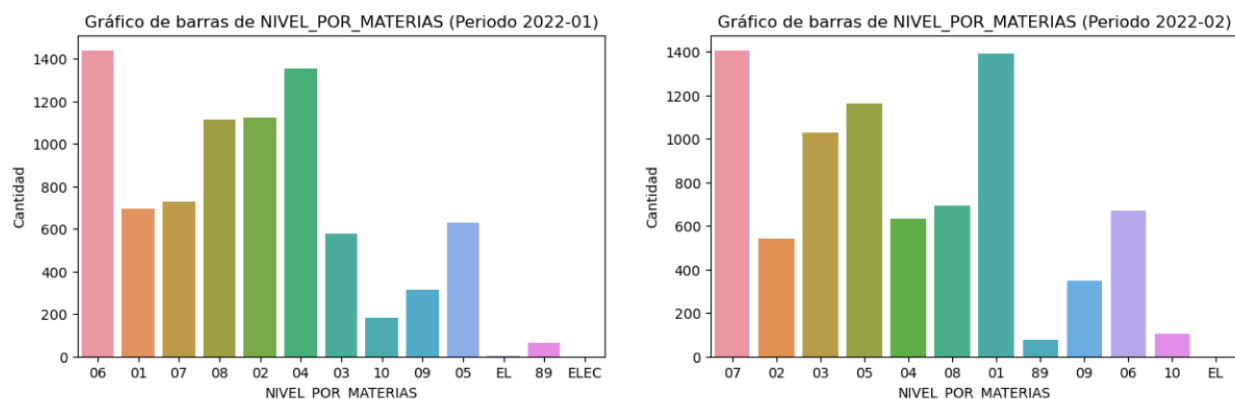


Ilustración 227

Gráfico de barras del tercer y cuarto periodo de la variable NIVEL_POR_MATERIAS del reporte de Becas



A partir del análisis llevado a cabo sobre las figuras exhibidas, se puede observar que el nivel en el que se encuentra un estudiante no presenta una relación directa con su aparición en el reporte de becas, ya que se observa una gran variabilidad entre los periodos. Esto tiene sentido, ya que un estudiante puede ser beneficiario de becas socioeconómicas en cualquier nivel académico, no existe un impedimento para recibir este tipo de beneficio en un nivel en particular, debido a que, la elegibilidad para las becas socioeconómicas no está determinada por el nivel en el que se encuentra el estudiante.

De manera similar, se realizará en este momento el análisis del nivel al que pertenece una asignatura aprobada por un estudiante en el reporte de becas, para esto, nuevamente se utilizará la técnica de gráficos de barras debido a la variedad de niveles existentes. El código, gráficos y conclusiones se pueden hallar a continuación:

Ilustración 228

Código que genera los gráficos para la variable NIVEL_POR_ASIG_APROBADAS del reporte de Becas

```
# Número de filas y columnas para los subplots
nrows = 2
ncols = 2

# Crear la figura y los ejes para los subplots
fig, axes = plt.subplots(nrows=nrows, ncols=ncols, figsize=(15, 10))

# Nombre de la columna de interés
columna_interes = 'NIVEL_POR_ASIG_APROBADAS'

# Iterar a través de los dataframes y crear un gráfico para cada uno
for idx, (beca, ax) in enumerate(zip(becas, axes.flat)):
    sns.countplot(x=columna_interes, data=beca, ax=ax)
    ax.set_xlabel(f"{columna_interes}")
    ax.set_ylabel('Cantidad')
    ax.set_title(f'Gráfico de barras de {columna_interes} ({get_periodo(beca.name)})')

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.4)

# Mostrar la figura
plt.show()
```

Ilustración 229

Gráfico de barras del primer y segundo periodo de la variable NIVEL_POR_ASIG_APROBADAS del reporte de Becas

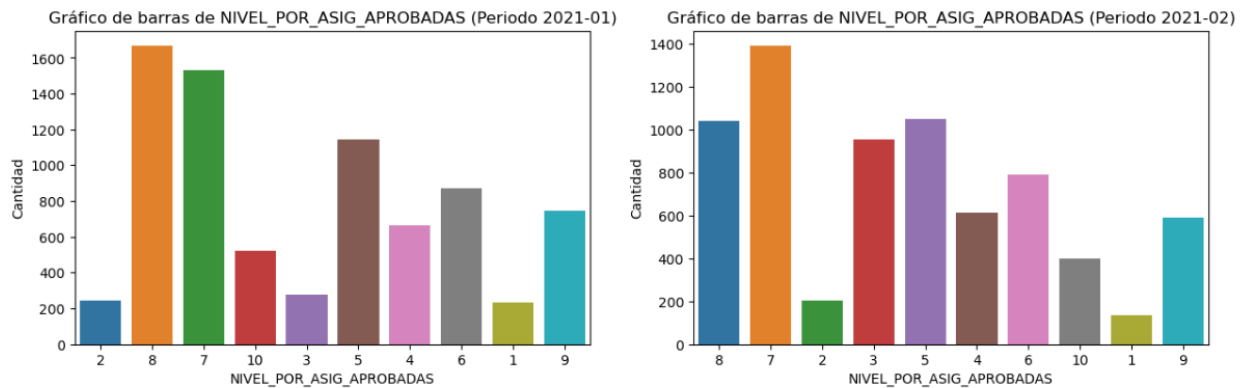
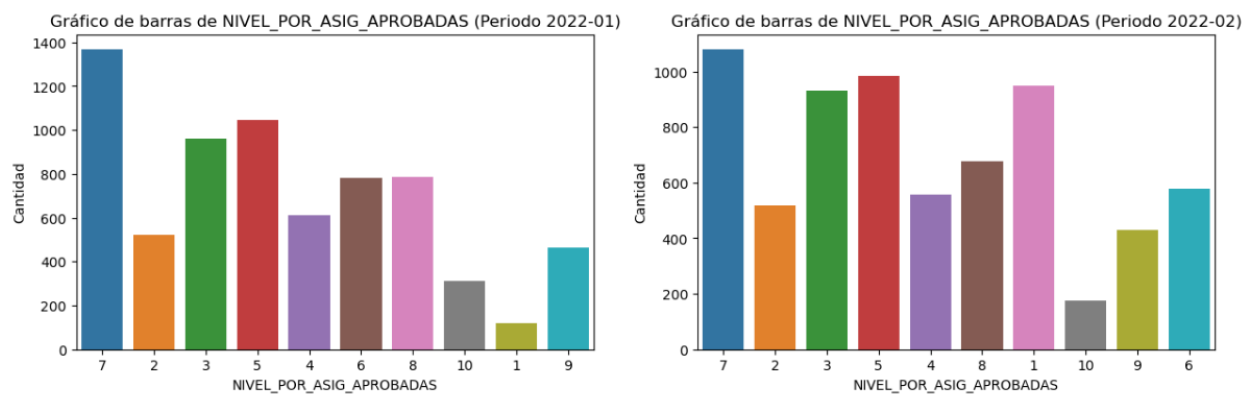


Ilustración 230

Gráfico de barras del tercer y cuarto periodo de la variable NIVEL_POR_ASIG_APROBADAS del reporte de Becas



A partir de un análisis sobre los gráficos de la variable, se puede concluir que a pesar de que pareciera no existir una relación directa entre el nivel de una asignatura aprobada y la acreditación de beca de un estudiante, también parece existir un cierto patrón de comportamiento por lo que se requiere un mayor análisis.

De manera diferente, ha llegado el momento de realizar un análisis avanzado de la variable correspondiente a que tipo de beca socioeconómica posee un determinado estudiante, para esto se utilizarán gráficos de barras ya que se desea visualizar los periodos con 0 registros

para que sea más intuitivo los análisis realizados sobre las estadísticas descriptivas. El código, gráficos y conclusiones serán mostradas a continuación:

Ilustración 231

Código que genera los gráficos para la variable CATEGORIA_BECA_SOCIOECONOMICA del reporte de Becas

```
# Número de filas y columnas para los subplots
nrows = 2
ncols = 2

# Crear la figura y los ejes para los subplots
fig, axes = plt.subplots(nrows=nrows, ncols=ncols, figsize=(15, 10))

# Nombre de la columna de interés
columna_interes = 'CATEGORIA_BECA_SOCIOECONOMICA'

# Iterar a través de los dataframes y crear un gráfico para cada uno
for idx, (beca, ax) in enumerate(zip(becas, axes.flat)):
    try:
        sns.countplot(x=columna_interes, data=beca, ax=ax)
    except:
        pass
    ax.set_xlabel(f"{columna_interes}")
    ax.set_ylabel('Cantidad')
    ax.set_title(f'Gráfico de barras de {columna_interes} ({get_periodo(beca.name)})')

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.4)

# Mostrar la figura
plt.show()
```

Ilustración 232

Gráfico de barras del primer y segundo periodo de la variable CATEGORIA_BECA_SOCIOECONOMICA del reporte de Becas

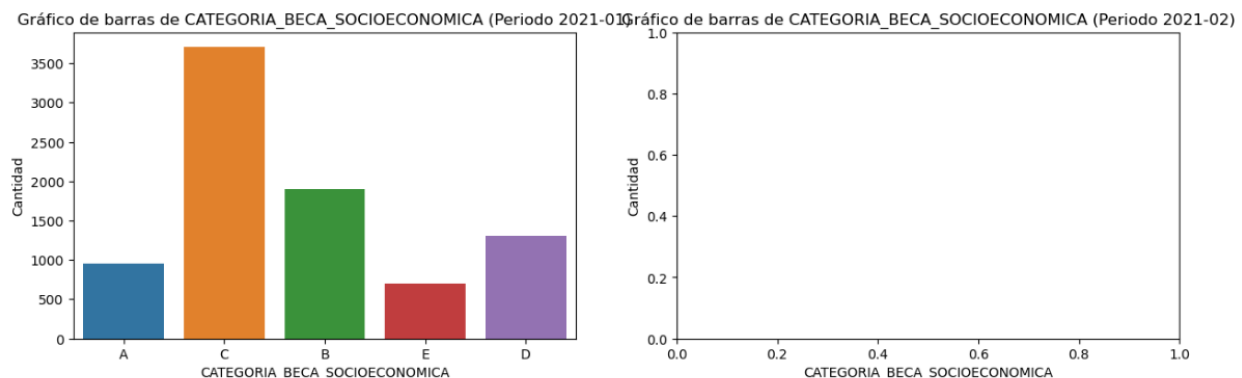
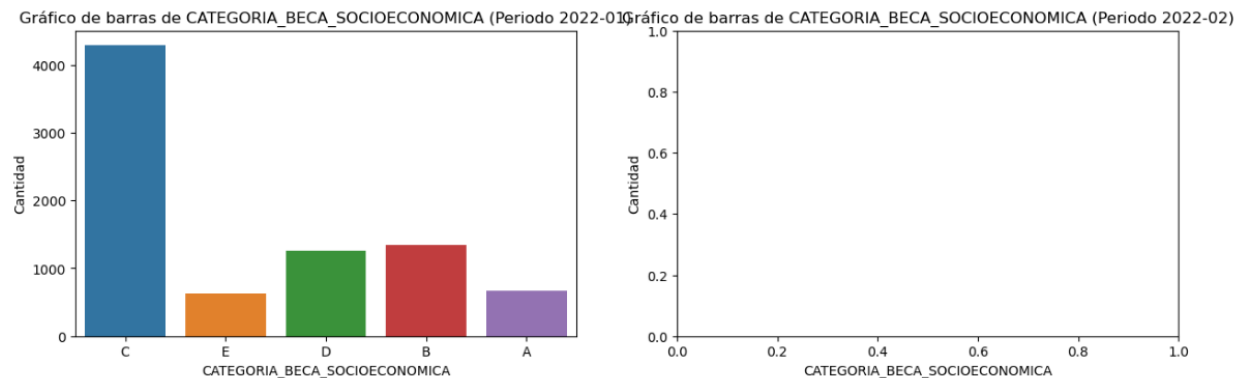


Ilustración 233

Gráfico de barras del tercer y cuarto periodo de la variable CATEGORIA_BECA_SOCIOECONOMICA del reporte de Becas



A partir de la visualización de los gráficos son evidentes las observaciones de las estadísticas descriptivas, además, se puede observar que los registros únicamente abarcan dos periodos académicos, por lo que, es recomendable realizar una revisión exhaustiva para identificar y manejar posibles duplicados asegurando la integridad de los datos antes de realizar análisis o conclusiones basadas en ellos.

En penúltimo término, se evaluará la variable TIENE_PAGO haciendo uso de gráficos de pastel debido a que únicamente se poseen dos tipos de datos. El código, gráfico y conclusiones se podrán encontrar a continuación:

Ilustración 234

Código que genera los gráficos para la variable TIENE_PAGO del reporte de Becas

```
# Se especifica el número de filas y columnas para los subplots
nrows = 2
ncols = 2

# Se crea la figura y los ejes para los subplots
fig, axes = plt.subplots(nrows=nrows, ncols=ncols, figsize=(15, 10))

# Se establece el nombre de la columna de interés
columna_interes = 'TIENE_PAGO'

# Se itera a través de los dataframes y se crea un gráfico de pastel para cada uno
for idx, (beca, ax) in enumerate(zip(becas, axes.flat)):
    try:
        # Se obtiene la cantidad de cada categoría en la columna de interés
        counts = beca[columna_interes].value_counts()
        # Se obtienen las etiquetas y tamaños para el gráfico de pastel
        labels = counts.index
        sizes = counts.values
        # Se crea el gráfico de pastel en los ejes actuales
        ax.pie(sizes, labels=labels, autopct='%1.1f%%', startangle=90)
    except:
        # En caso de que no se pueda generar un gráfico de pastel, se omite y se pasa al siguiente
        pass

    # Se establece el título para cada subplot
    ax.set_title(f'Gráfico de pastel de {columna_interes} ({get_periodo(beca.name)})')

# Se ajusta el espacio entre los subplots para evitar superposiciones
plt.subplots_adjust(wspace=0.2, hspace=0.4)

# Se muestra la figura con los subplots
plt.show()
```

Ilustración 235

Gráfico de barras del primer y segundo periodo de la variable TIENE_PAGO del reporte de Becas

Gráfico de pastel de TIENE_PAGO (Periodo 2021-01)

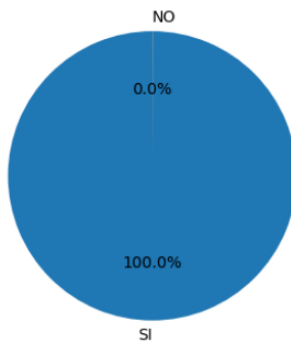


Gráfico de pastel de TIENE_PAGO (Periodo 2021-02)

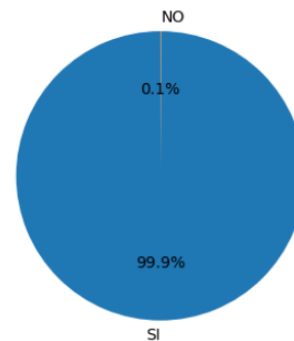


Ilustración 236

Gráfico de barras del tercer y cuarto periodo de la variable TIENE_PAGO del reporte de Becas

Gráfico de pastel de TIENE_PAGO (Periodo 2022-01)

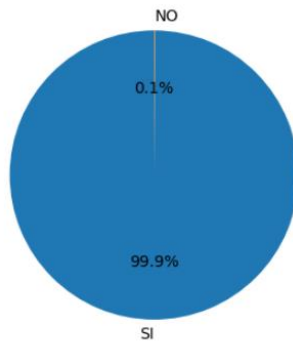
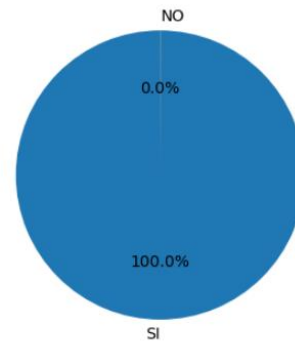


Gráfico de pastel de TIENE_PAGO (Periodo 2022-02)



A partir de la simple observaciones de los gráficos se puede determinar la inmensa desproporción en tipos de datos de la variable, por lo que, dado que puede generar un sesgo en el modelo de machine learning planteado por esta investigación y desencadenar un posible sobreentrenamiento, lo más prudente sería cesar el uso de esta variable.

De manera similar, el análisis de la variable TIENE_SALDO se determinará a partir de gráficos de pastel por las mismas razones que la variable TIENE_PAGO. El código, gráficos y conclusiones serán exhibidas a continuación:

Ilustración 237

Código que genera los gráficos para la variable TIENE_SALDO del reporte de Becas

```
# Se especifica el número de filas y columnas para los subplots
nrows = 2
ncols = 2

# Se crea la figura y los ejes de los subplots con el tamaño especificado
fig, axes = plt.subplots(nrows=nrows, ncols=ncols, figsize=(15, 10))

# Se establece el nombre de la columna de interés
columna_interes = 'TIENE_SALDO'

# Colores para los gráficos de pie
colores = ['#FF7F50', '#87CEEB']

# Se itera a través de los dataframes y se crea un gráfico de pie para cada uno
for idx, (beca, ax) in enumerate(zip(becas, axes.flat)):
    try:
        # Se obtiene la cantidad de cada categoría en la columna de interés
        counts = beca[columna_interes].value_counts()
        # Se obtienen las etiquetas y tamaños para el gráfico de pie
        labels = counts.index
        sizes = counts.values
        # Se crea el gráfico de pastel en los ejes actuales
        ax.pie(sizes, labels=labels, autopct='%1.1f%%', startangle=90, colors=colores)
    except:
        # En caso de que no se pueda generar un gráfico de pie, se omite y se pasa al siguiente
        pass

    # Se establece el título para cada subplot
    ax.set_title(f'Gráfico de pastel de {columna_interes} ({get_periodo(beca.name)})')

# Se ajusta el espacio entre los subplots para evitar superposiciones
plt.subplots_adjust(wspace=0.2, hspace=0.4)

# Se muestra la figura con los subplots
plt.show()
```

Ilustración 238

Gráfico de barras del primer y segundo periodo de la variable TIENE_SALDO del reporte de Becas

Gráfico de pastel de TIENE_SALDO (Periodo 2021-01)

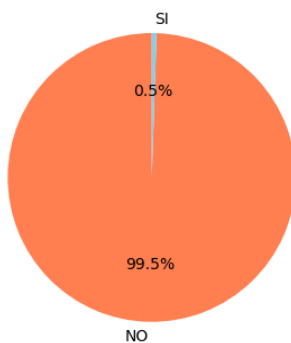


Gráfico de pastel de TIENE_SALDO (Periodo 2021-02)

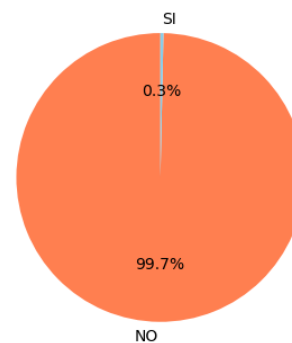


Ilustración 239

Gráfico de barras del tercer y cuarto periodo de la variable TIENE_SALDO del reporte de Becas

Gráfico de pastel de TIENE_SALDO (Periodo 2022-01)

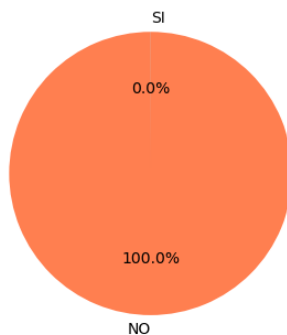
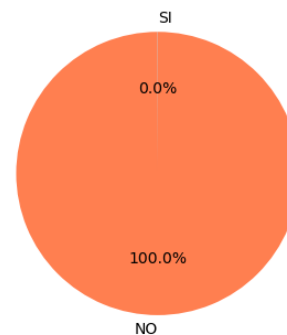


Gráfico de pastel de TIENE_SALDO (Periodo 2022-02)



Como último análisis de variables categóricas, a partir de los gráficos de pastel se puede determinar que en todos los periodos existe un número mínimo de estudiantes que no realizan el pago del periodo académico, por lo que, genera una abismal desproporción de datos. Es decir, así como la variable TIENE_PAGO, la acción con mayor prudencia es el cesar del uso de la variable TIENE_SALDO.

En conclusión, como acto final de la exploración de variables categóricas en el reporte de becas, se ha logrado con éxito identificar aquellas variables que, sin importar que no se haya realizado una exploración de relaciones con ellas, no aportarán nada a la investigación, pues, de hecho, podrían incluso perjudicar al modelo de predicción de rendimiento académico, estas son:

- NACIONALIDAD
- AREA_MATERIAS
- TIENE_PAGO
- TIENE_SALDO

4.1.12. Exploración de relaciones

En este capítulo, el objetivo principal es identificar y examinar las relaciones potenciales que pueden existir entre las diferentes variables en estudio. (Alteryx, s. f.). Por lo que, para cumplir este objetivo, se debe determinar en primer lugar una variable de respuesta ante la búsqueda de una relación (Orellana, 2001). Es decir, para lograr enfocar una correcta búsqueda de relaciones entre las variables que se disponen, es necesario especificar aquellas que serán usadas para validar una relación, en otras palabras, correspondiendo al contexto de la investigación presente, es necesario definir hipótesis que validarán la relación que existe entre una variable y el rendimiento académico.

Adicionalmente, para realizar dicha exploración, se recurrirá al uso de gráficos, entre ellos, el conocido diagrama de violín que consiste en un diagrama de caja y bigotes fusionado con el diagrama de densidad kernel, esto, permitirá indicar en que puntos los datos muestran una mayor distribución. (Carron, 2021). Por lo que, a partir de uso y de diagramas de cajas y bigotes, dada la necesidad y la disposición de los datos de las variables, se intentará detectar patrones, correlaciones o tendencias que puedan aportar una visión más profunda de la investigación (jmp, s. f.) Y, principalmente, responder a las hipótesis planteadas.

Sin embargo, cabe resaltar que el objetivo de esta exploración no es confirmar hipótesis estadísticas (jmp, s. f.). Pues, como se mencionó al comienzo, se debe responder hipótesis que permitan identificar las relaciones entre variables dado el contexto de la investigación; por lo que, este análisis exploratorio es esencial no solo para entender mejor los datos, sino también para formular una hipótesis estructurada para el planteamiento del modelo de machine learning de la predicción del rendimiento académico.

Entonces, a partir de lo conversado hasta este momento, se debe apreciar que la exploración de las relaciones será enfocada hacia el rendimiento académico, es decir, se analizarán escenarios donde existe una posibilidad de que se afecte al rendimiento académico; por lo que la evaluación y análisis de relaciones entre variables inocuas al rendimiento académico no se realizarán, como, por ejemplo, un análisis entre las facultades y el género, pues si bien el comprender la distribución de estudiantes dado el género puede ayudar a generar programas para apoyar y motivar a ciertos segmentos poblacionales a estudiar carreras pertenecientes a estas facultades, esta área no compete a la investigación que se está realizando.

Por lo que las hipótesis para el modelo de rendimiento académico que se tratarán en este subcapítulo de descubrimiento de relaciones serán detalladas en los siguientes contextos:

En el contexto del reporte de asignaturas perdidas

En primer lugar, hay que establecer que variables reflejar el rendimiento académico, estas son:

- **NOTA_FINAL:** Refleja la calificación final del estudiante en la asignatura.
- **VECES:** Refleja la cantidad de veces que el estudiante ha repetido la asignatura.
- **APROBADO:** Refleja si el estudiante ha aprobado o no una asignatura (Al ser todos los registros del reporte asignaturas reprobadas, no se va a generar una información útil para ningún análisis).

Una vez establecido que variables reflejan el rendimiento académico de un determinado estudiante en una determinada asignatura, es necesario mencionar al ser todos los registros del reporte de asignaturas perdidas correspondientes a estudiantes reprobados y, por tanto, ningún registro de estudiantes aprobados, una hipótesis sobre el campo APROBADO sería irrelevante;

Por lo que, se puede plantear las siguientes hipótesis con respecto al rendimiento académico para la exploración de relaciones:

- Las variables referidas al plan de estudios del estudiante, su facultad, carrera, nivel, créditos de la asignatura y área de la asignatura influyen en el rendimiento de la nota final de los estudiantes.
- Las variables referidas al plan de estudios del estudiante, su facultad, carrera, nivel, créditos de la asignatura y área de la asignatura influyen en la cantidad de veces que el estudiante ha repetido alguna asignatura.

En el contexto del reporte de calificaciones por carrera

A igual que en el segmento anterior, como acto inicial se debe identificar las variables que reflejan el rendimiento académico, estas son:

- `nota_final`: Refleja la calificación final del estudiante en la asignatura

Una vez identificado que la variable `nota_final` es la única que refleja directamente el rendimiento académico de un determinado estudiante en una determinada asignatura, se puede plantear la siguiente hipótesis para el modelo de machine learning objetivo de este trabajo de titulación:

- Las variables referidas a la facultad, docente, área de la asignatura y carrera influyen en el rendimiento de las notas finales de un estudiante.

En el contexto del reporte de becas

Con respecto al último segmento, de manera similar, en primera instancia es necesario identificar las variables que reflejan el rendimiento académico, estas son:

- NUM_MATERIAS_SEGUNDA_MATRICULA: Refleja la cantidad de asignaturas con segunda matrícula que posee un estudiante.
- NUM_MATERIAS_TERCERA_MATRICULA: Refleja la cantidad de asignaturas con tercera matrícula que posee un estudiante.

Una vez establecido que estas variables reflejan el rendimiento académico de un determinado estudiante, es permisible el planteamiento de las siguientes hipótesis para el modelo de machine learning de predicción del rendimiento académico:

- Las variables referidas al género, etnia, tipo de colegio, carrera, nivel del estudiante, nivel de asignaturas tomadas, categoría de beca socioeconómica influyen en el rendimiento académico con respecto a la cantidad de materias en segunda matrícula.
- Las variables referidas al género, etnia, tipo de colegio, carrera, nivel del estudiante, nivel de asignaturas tomadas, categoría de beca socioeconómica y cantidad de asignaturas con segunda matrícula influyen en el rendimiento académico con respecto a la cantidad de materias en tercera matrícula.

En resumen, hasta el momento, se ha logrado detectar las variables que reflejan el rendimiento académico y se han planteado hipótesis a resolver para la exploración de relaciones de las variables y el rendimiento académico, por lo que, a continuación, se realizarán propiamente las exploraciones, inicialmente por el reporte de asignaturas perdidas.

Por lo tanto, el análisis de exploración de relaciones comenzará con la evaluación de la relación entre los códigos del plan de estudios de un estudiante y su nota final haciendo uso del diagrama de violín. El código, gráficos y conclusiones se podrán observar a continuación:

Ilustración 240

Código que genera los gráficos para la relación entre PLAN_E y NOTA_FINAL del reporte de Asignaturas Perdidas

```
# Ordenamos las categorías alfabéticamente
orden_categorias = set()
for perdida in perdidas:
    orden_categorias.update(list(sorted(perdida["PLAN_E"].unique())))

# Grid para los gráficos
fig, axes = plt.subplots(4, 1, figsize=(10, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='PLAN_E', y='NOTA_FINAL', data=perdidas[0], order=orden_categorias, ax=axes[0])

axes[1].set_title('Periodo 2021-02')
sns.violinplot(x='PLAN_E', y='NOTA_FINAL', data=perdidas[1], order=orden_categorias, ax=axes[1])

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='PLAN_E', y='NOTA_FINAL', data=perdidas[2], order=orden_categorias, ax=axes[2])

axes[3].set_title('Periodo 2022-02')
sns.violinplot(x='PLAN_E', y='NOTA_FINAL', data=perdidas[3], order=orden_categorias, ax=axes[3])

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.4)

plt.show() # Mostrar los gráficos
```

Ilustración 241

Diagrama de Violín del primer periodo de la relación entre PLAN_E y NOTA_FINAL del reporte de Asignaturas Perdidas

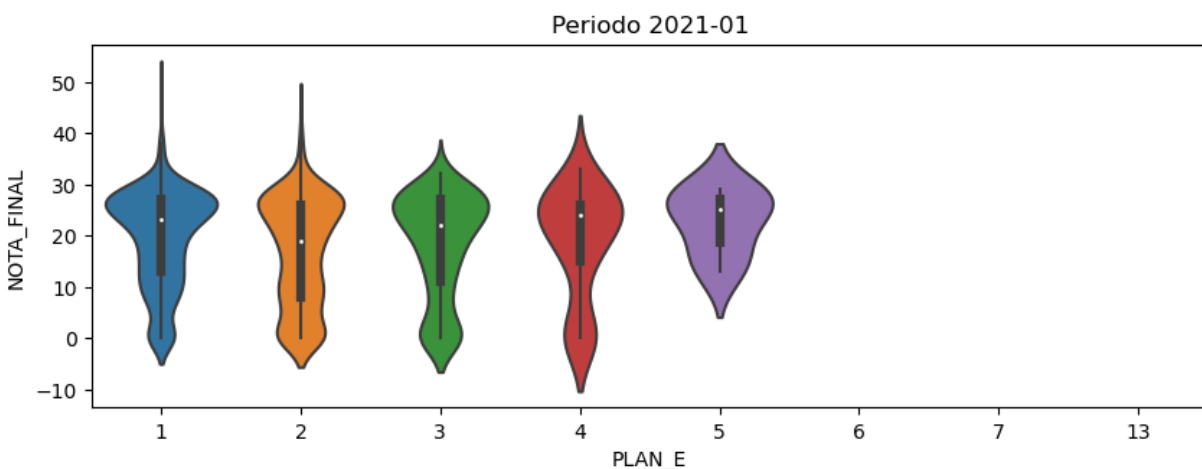


Ilustración 242

Diagrama de Violín del segundo periodo de la relación entre PLAN_E y NOTA_FINAL del reporte de Asignaturas Perdidas

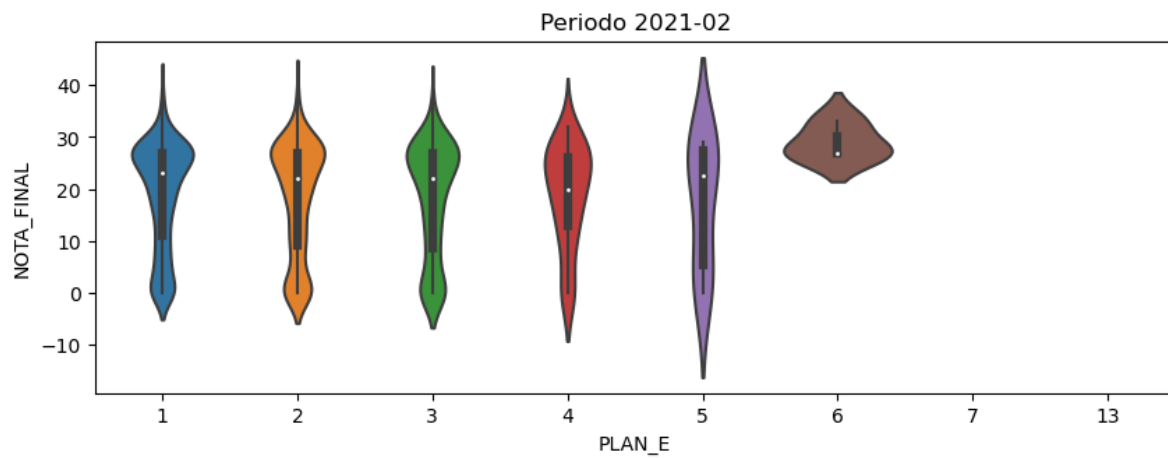


Ilustración 243

Diagrama de Violín del tercer periodo de la relación entre PLAN_E y NOTA_FINAL del reporte de Asignaturas Perdidas

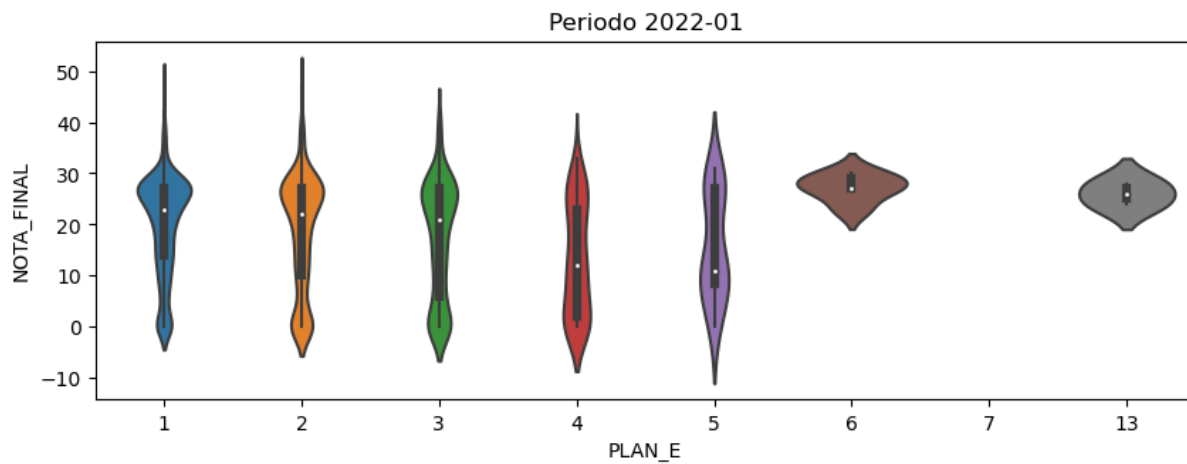
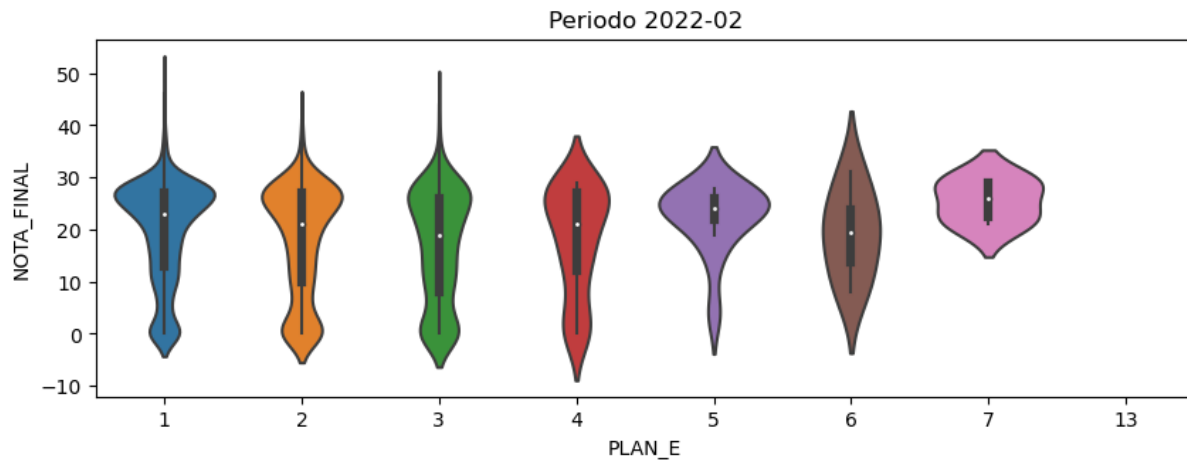


Ilustración 244

Diagrama de Violín del cuarto periodo de la relación entre PLAN_E y NOTA_FINAL del reporte de Asignaturas Perdidas



A partir de los diagramas se puede concluir que la distribución de las relaciones varía considerablemente de un periodo académico a otro, adicionalmente, existe mucha similitud entre las diferentes distribuciones de los planes de estudio, esto podría simbolizar que no existe una correlación entre el plan de estudio y el rendimiento académico. En otras palabras, aunque existen diferencias en las distribuciones de calificaciones finales entre cada periodo, estas diferencias no parecen estar relacionadas con el plan de estudio que los estudiantes están siguiendo, pues, las similitudes en las distribuciones de calificaciones entre los diferentes planes de estudio se mantienen constantes; debido a esto, se refuerza la idea de que no hay una correlación evidente entre el rendimiento final de los estudiantes y su plan de estudio.

Por lo tanto, es plausible concluir que, a pesar de la variabilidad en las calificaciones finales entre diferentes periodos y ciertas diferencias mínimas entre planes, el plan de estudio no parece tener un impacto significativo en estas calificaciones. No obstante, cabe recalcar que en algunos casos se ve más marcado un cambio significativo entre planes de estudio, pero, al ser

planes que no poseen registros en algunos periodos como por ejemplo el plan 7 que únicamente lo posee el último periodo, no representa utilidad y podría afectar a la precisión del modelo.

Por otra parte, con respecto a la relación las facultades a las que pertenecen los estudiantes y sus notas finales se realizarán utilizando un diagrama de violín cuyo respectivo código, gráficos y conclusiones se podrán encontrar a continuación:

Ilustración 245

Código que genera los gráficos para la relación entre DESC_FACULTAD y NOTA_FINAL del reporte de Asignaturas Perdidas

```
# Ordenamos las categorías alfabéticamente
orden_categorias = set()
for perdida in perdidas:
    orden_categorias.update(list(sorted(perdida["DESC_FACULTAD"].unique())))

# Grid para los gráficos
fig, axes = plt.subplots(4, 1, figsize=(10, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='DESC_FACULTAD', y='NOTA_FINAL', data=perdidas[0], order=orden_categorias, ax=axes[0])
axes[0].set_xticklabels(axes[0].get_xticklabels(), rotation=90)

axes[1].set_title('Periodo 2021-02')
sns.violinplot(x='DESC_FACULTAD', y='NOTA_FINAL', data=perdidas[1], order=orden_categorias, ax=axes[1])
axes[1].set_xticklabels(axes[1].get_xticklabels(), rotation=90)

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='DESC_FACULTAD', y='NOTA_FINAL', data=perdidas[2], order=orden_categorias, ax=axes[2])
axes[2].set_xticklabels(axes[2].get_xticklabels(), rotation=90)

axes[3].set_title('Periodo 2022-02')
sns.violinplot(x='DESC_FACULTAD', y='NOTA_FINAL', data=perdidas[3], order=orden_categorias, ax=axes[3])
axes[3].set_xticklabels(axes[3].get_xticklabels(), rotation=90)

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=2.4)

plt.show() # Mostrar los gráficos
```

Ilustración 246

Diagrama de Violín del primer periodo de la relación entre DESC_FACULTAD y NOTA_FINAL del reporte de Asignaturas Perdidas

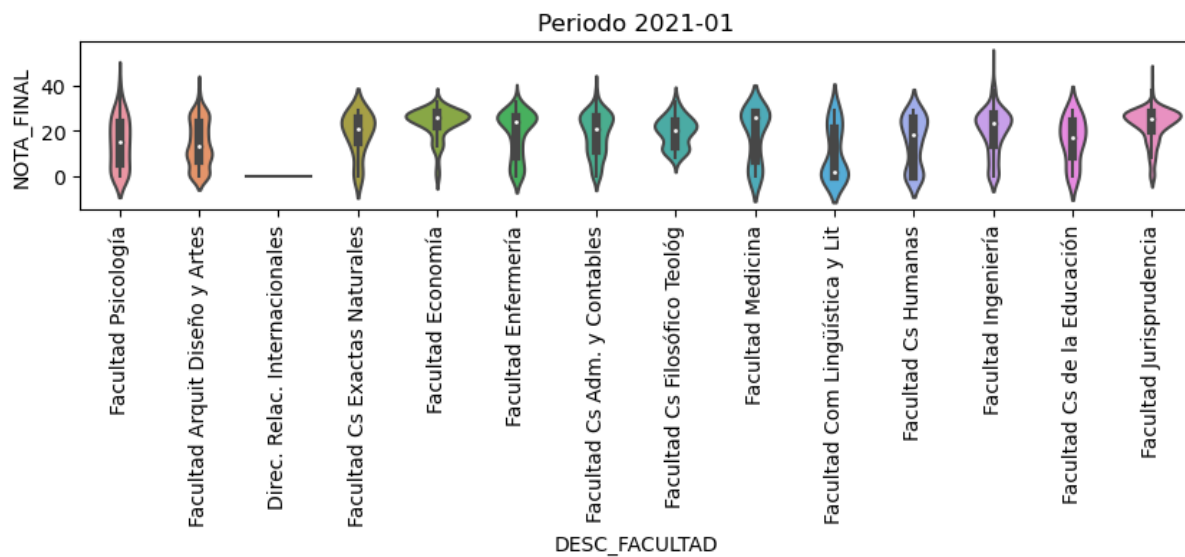


Ilustración 247

Diagrama de Violín del segundo periodo de la relación entre DESC_FACULTAD y NOTA_FINAL del reporte de Asignaturas Perdidas

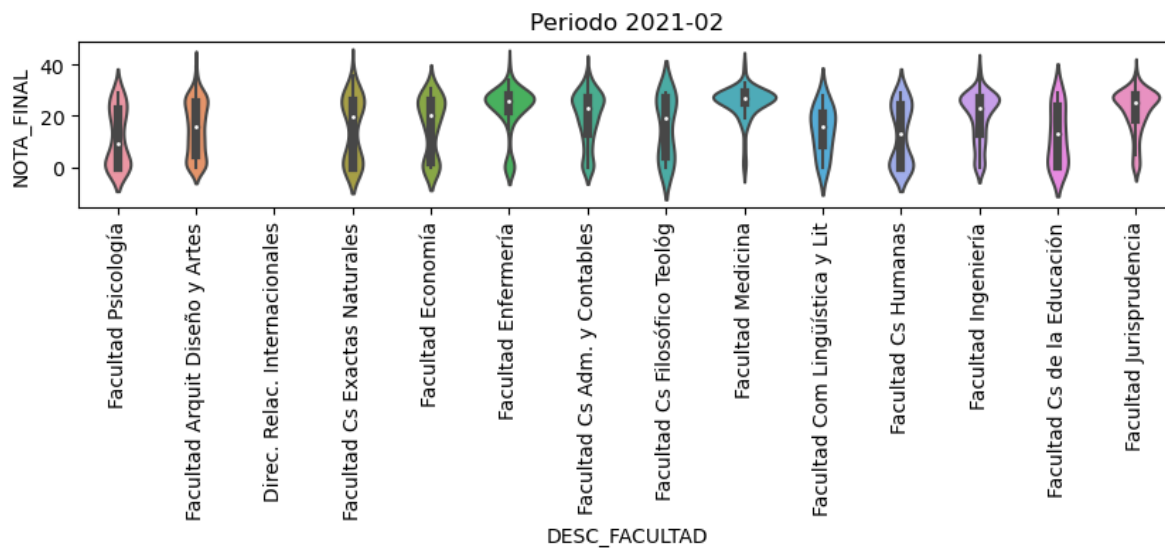


Ilustración 248

Diagrama de Violín del tercer periodo de la relación entre DESC_FACULTAD y NOTA_FINAL del reporte de Asignaturas Perdidas

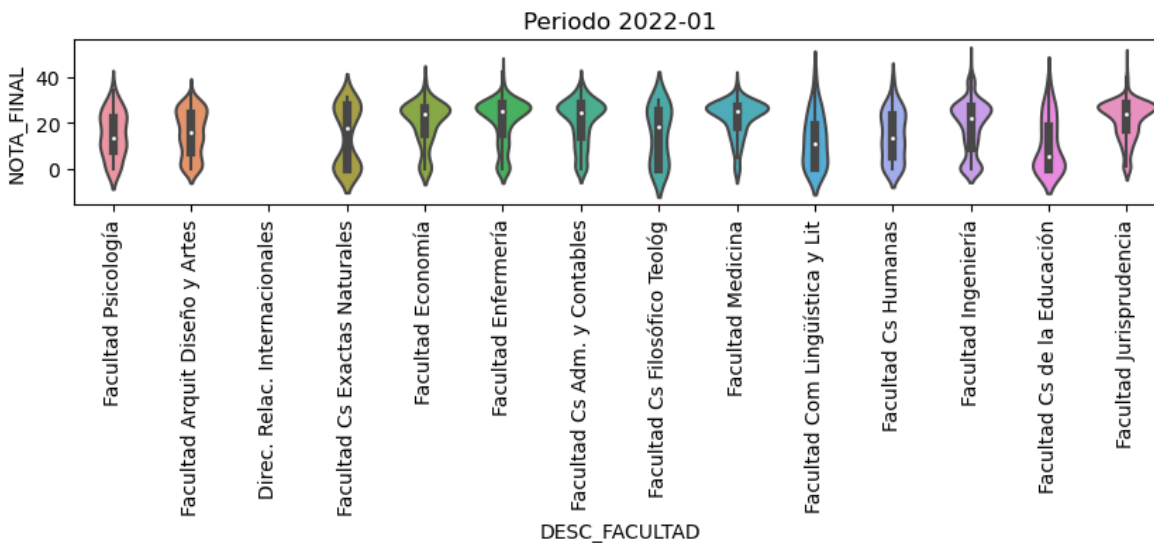
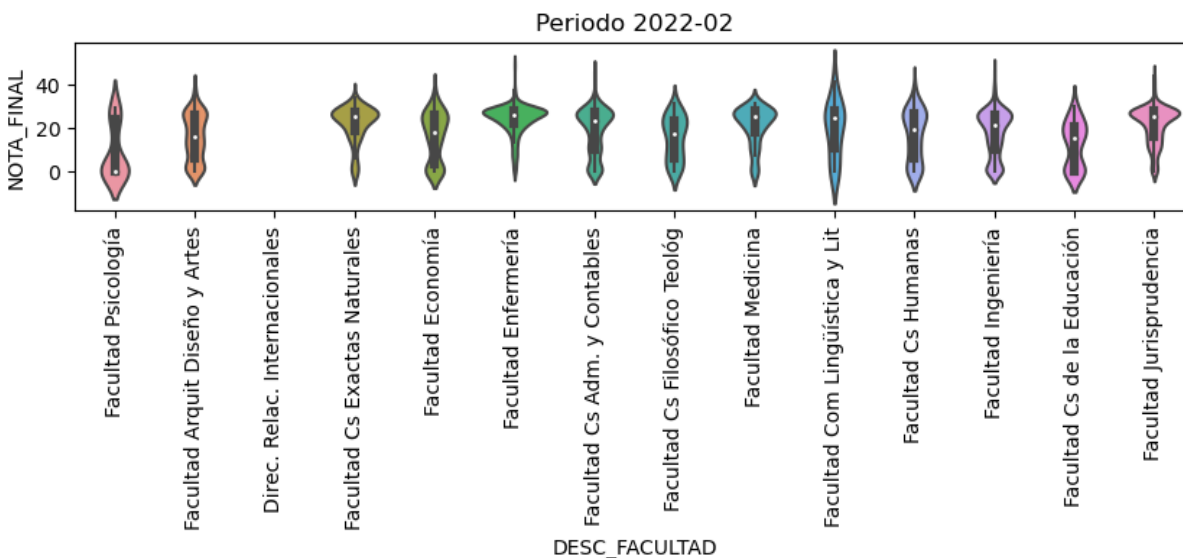


Ilustración 249

Diagrama de Violín del cuarto periodo de la relación entre DESC_FACULTAD y NOTA_FINAL del reporte de Asignaturas Perdidas



A partir del análisis realizado sobre la variable que representa la facultad de los estudiantes, se pudo concluir que, si bien algunas facultades muestran una variabilidad considerable en la distribución de las notas finales, es evidente que todas las distribuciones

siguen un cierto patrón, lo que sugiere que podría existir una relación directa entre la facultad en la que un estudiante se encuentra y su desempeño académico. En otras palabras, las diferencias en las distribuciones de notas finales entre las facultades pueden indicar que cada facultad tiene características específicas que influyen en el rendimiento de los estudiantes, dichas características pueden incluir el nivel de dificultad de los programas de estudio, el enfoque pedagógico utilizado, los recursos disponibles o incluso la cultura académica en cada facultad. Por lo que, se puede concluir que esta variable es importante para el modelo de machine learning deseado.

En otro sentido, para considerar la relación entre las carreras universitarias de los estudiantes y sus notas finales, se utilizará un diagrama de caja y bigotes. El código, figuras y conclusiones se podrán observar a continuación:

Ilustración 250

Código que genera los gráficos para la relación entre DESC_PROGRAMA y NOTA_FINAL del reporte de Asignaturas Perdidas

```
# Ordenamos las categorías alfabéticamente
orden_categorias = set()
for perdida in perdidas:
    orden_categorias.update(list(sorted(perdida["DESC_PROGRAMA"].unique())))

# Grid para los gráficos
fig, axes = plt.subplots(4, 1, figsize=(12, 25))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.boxplot(x='DESC_PROGRAMA', y='NOTA_FINAL', data=perdidas[0], order=orden_categorias, ax=axes[0])
axes[0].set_xticklabels(axes[0].get_xticklabels(), rotation=90)

axes[1].set_title('Periodo 2021-02')
sns.boxplot(x='DESC_PROGRAMA', y='NOTA_FINAL', data=perdidas[1], order=orden_categorias, ax=axes[1])
axes[1].set_xticklabels(axes[1].get_xticklabels(), rotation=90)

axes[2].set_title('Periodo 2022-01')
sns.boxplot(x='DESC_PROGRAMA', y='NOTA_FINAL', data=perdidas[2], order=orden_categorias, ax=axes[2])
axes[2].set_xticklabels(axes[2].get_xticklabels(), rotation=90)

axes[3].set_title('Periodo 2022-02')
sns.boxplot(x='DESC_PROGRAMA', y='NOTA_FINAL', data=perdidas[3], order=orden_categorias, ax=axes[3])
axes[3].set_xticklabels(axes[3].get_xticklabels(), rotation=90)

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=1.4)

plt.show() # Mostrar los gráficos
```

Ilustración 251

Diagrama de Violín del primer periodo de la relación entre DESC_PROGRAMA y NOTA_FINAL del reporte de Asignaturas Perdidas

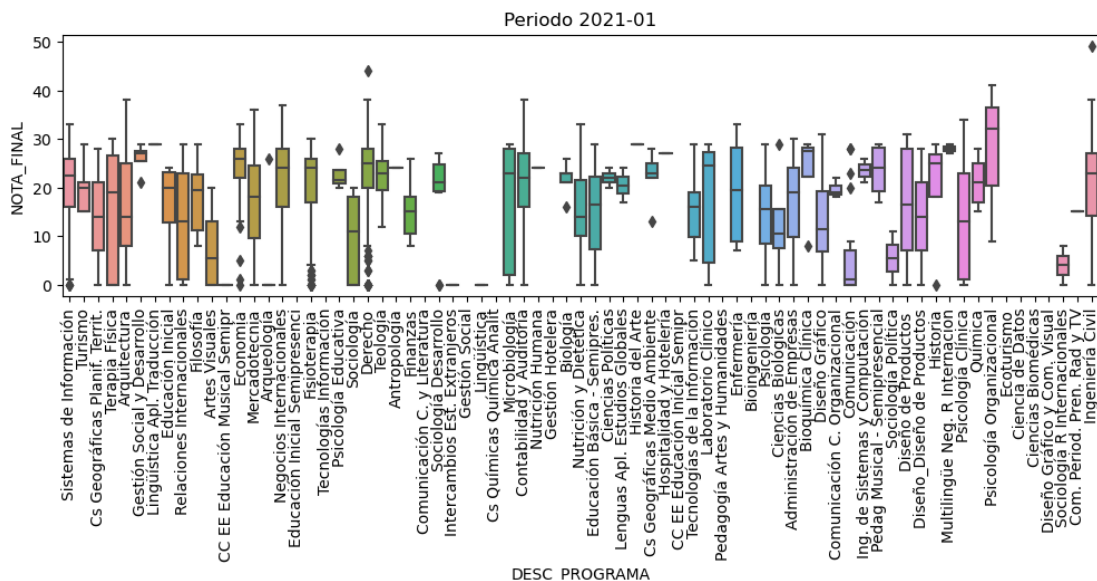


Ilustración 252

Diagrama de Violín del segundo periodo de la relación entre DESC_PROGRAMA y NOTA_FINAL del reporte de Asignaturas Perdidas

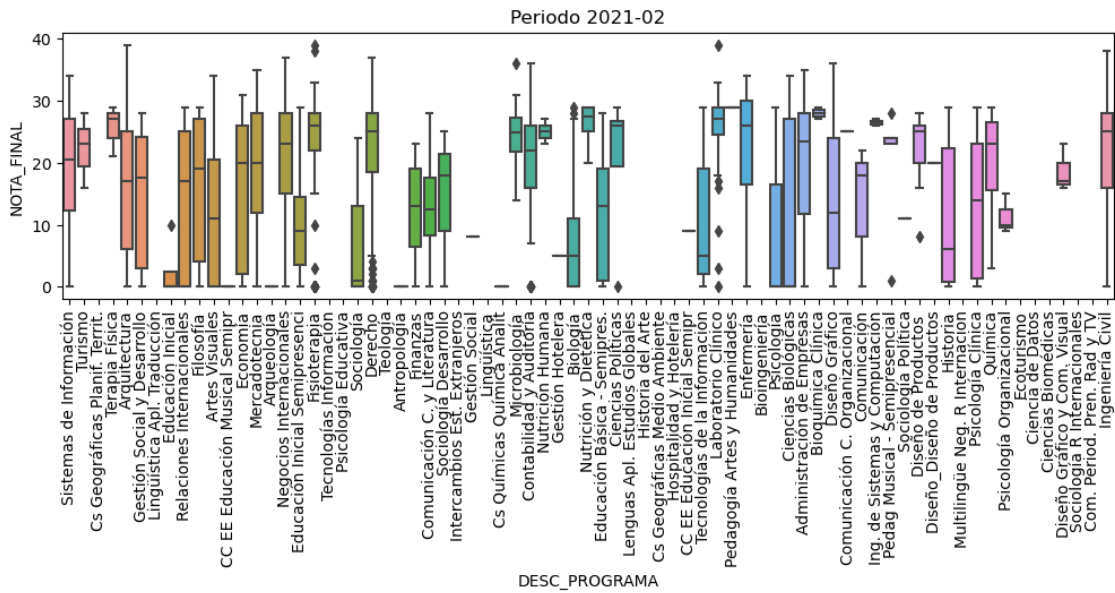


Ilustración 253

Diagrama de Violín del tercer periodo de la relación entre DESC_PROGRAMA y NOTA_FINAL del reporte de Asignaturas Perdidas

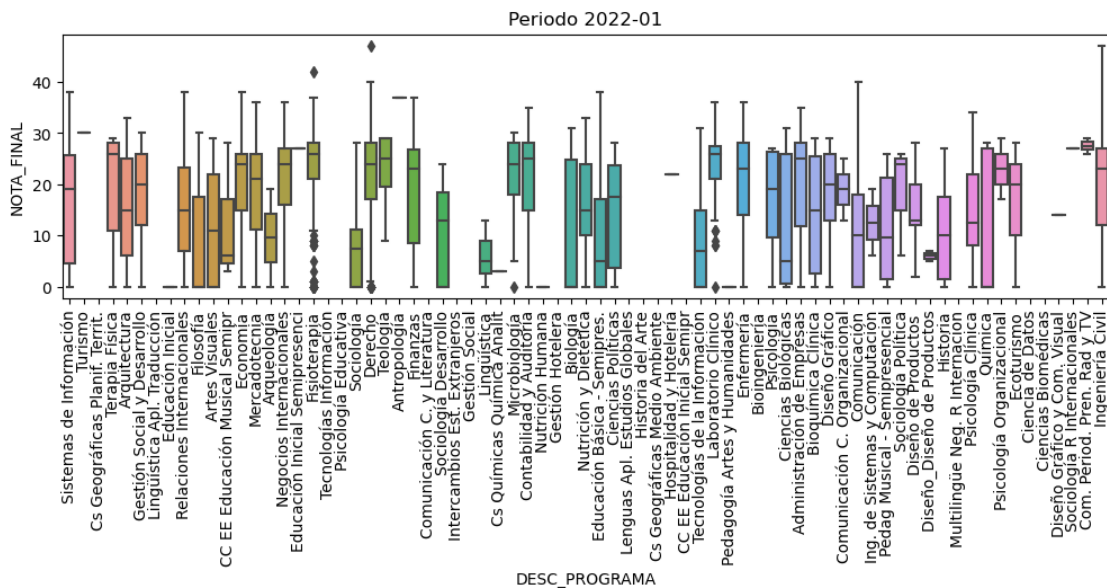
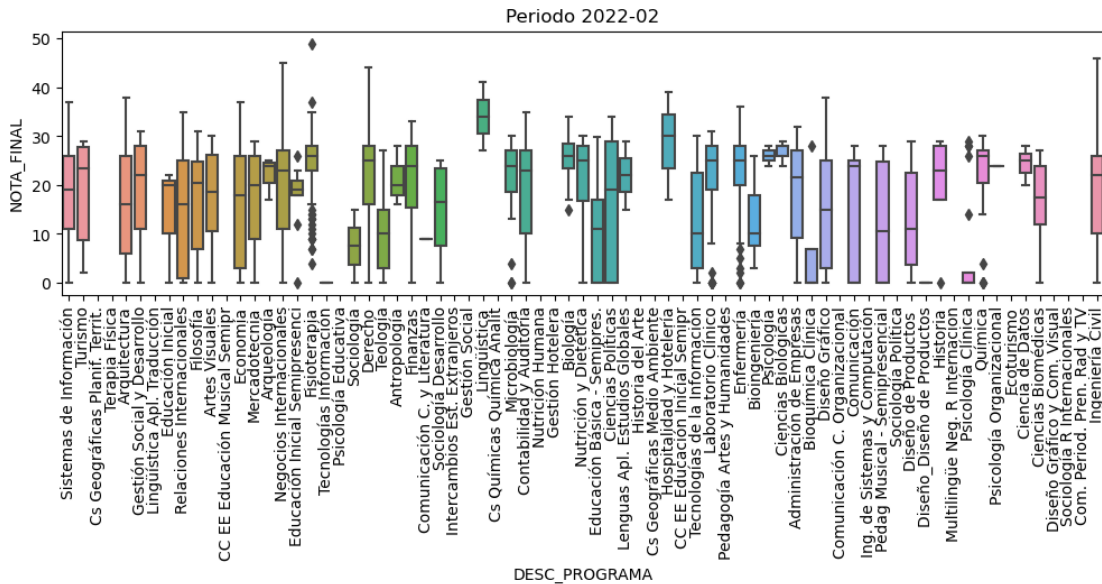


Ilustración 254

Diagrama de Violín del cuarto periodo de la relación entre DESC_PROGRAMA y NOTA_FINAL del reporte de Asignaturas Perdidas



Se puede concluir que la distribución de las calificaciones varía dependiendo de la carrera, sin embargo, en algunas carreras se puede observar un comportamiento similar a través de los periodos académicos; esto podría implicar que para el modelo de machine learning no se va a requerir de todas las carreras dispuestas en este conjunto de datos, ya que, se debe considerar que existe mucha variación en la distribución de calificaciones finales en algunas carreras, esto, da a entender que no existe una estrecha relación entre pertenecer a estas determinadas carreras y el rendimiento académico de un determinado estudiante.

Por otra parte, las carreras que, si muestran una estrecha relación con el rendimiento académico al no presentar variación considerable entre periodos académicos, sugiere que el rendimiento académico de un determinado estudiante puede verse influenciado por su asistencia a estas determinadas carreras; estas carreras son:

- Contabilidad y auditoría

- Arquitectura
- Diseño de productos
- Artes visuales
- Sistemas de información
- Mercadotecnia
- Derecho
- Finanzas
- Administración de empresas
- Tecnologías de la información
- Diseño Gráfico
- Enfermería
- Relaciones internacionales
- Educación básica semipresencial -> Educación Básica - Semipres.
- Microbiología
- Nutrición y Dietética
- Negocios Internacionales
- Ingeniería Civil
- Comunicación
- Filosofía
- Sociología
- Fisioterapia

Adicionalmente, según el análisis realizado, se puede inferir que las carreras en las que se observa una relación con el rendimiento académico podrían deberse a que implementan procesos

que no han sido modificados a lo largo del tiempo, lo cual puede afectar directamente la experiencia de los estudiantes en su vida universitaria, es decir, el hecho de que el desempeño académico esté relacionado con algunas carreras podría indicar un punto de alerta en el cual la manera en que se maneja la carrera afecta directamente el rendimiento de un estudiante.

Por otra parte, es momento de considerar la relación entre los niveles de los estudiantes y sus notas finales, la misma será resultada aplicando el diagrama de violín:

Ilustración 255

Código que genera los gráficos para la relación entre NIVEL_ESTUDIANTE y NOTA_FINAL del reporte de Asignaturas Perdidas

```
# Ordenamos las categorías alfabéticamente
orden_categorias = set()
for perdida in perdidas:
    orden_categorias.update(list(sorted(perdida["NIVEL_ESTUDIANTE"].unique())))

# Grid para los gráficos
fig, axes = plt.subplots(4, 1, figsize=(10, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='NIVEL_ESTUDIANTE', y='NOTA_FINAL', data=perdidas[0], order=orden_categorias, ax=axes[0])

axes[1].set_title('Periodo 2021-02')
sns.violinplot(x='NIVEL_ESTUDIANTE', y='NOTA_FINAL', data=perdidas[1], order=orden_categorias, ax=axes[1])

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='NIVEL_ESTUDIANTE', y='NOTA_FINAL', data=perdidas[2], order=orden_categorias, ax=axes[2])

axes[3].set_title('Periodo 2022-02')
sns.violinplot(x='NIVEL_ESTUDIANTE', y='NOTA_FINAL', data=perdidas[3], order=orden_categorias, ax=axes[3])

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.4)

plt.show() # Mostrar los gráficos
```

Ilustración 256

Diagrama de Violín del primer periodo de la relación entre NIVEL_ESTUDIANTE y NOTA_FINAL del reporte de Asignaturas Perdidas

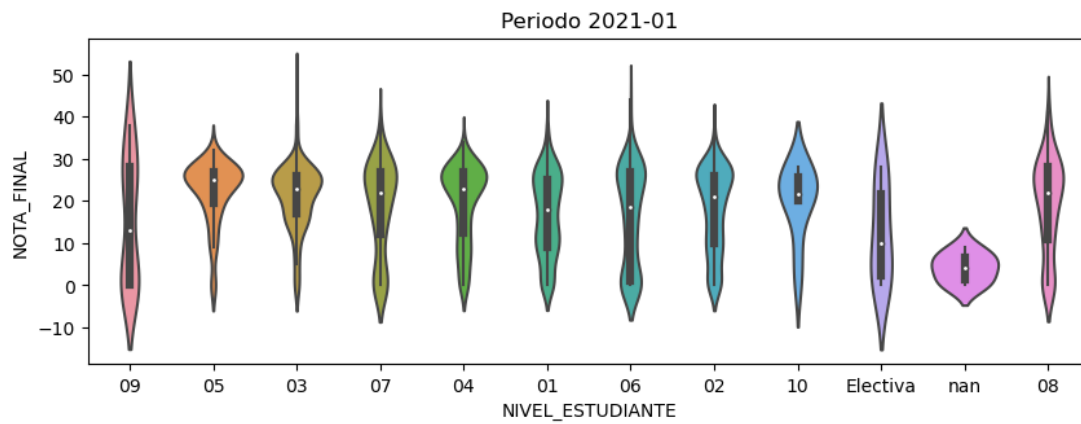


Ilustración 257

Diagrama de Violín del segundo periodo de la relación entre NIVEL_ESTUDIANTE y NOTA_FINAL del reporte de Asignaturas Perdidas

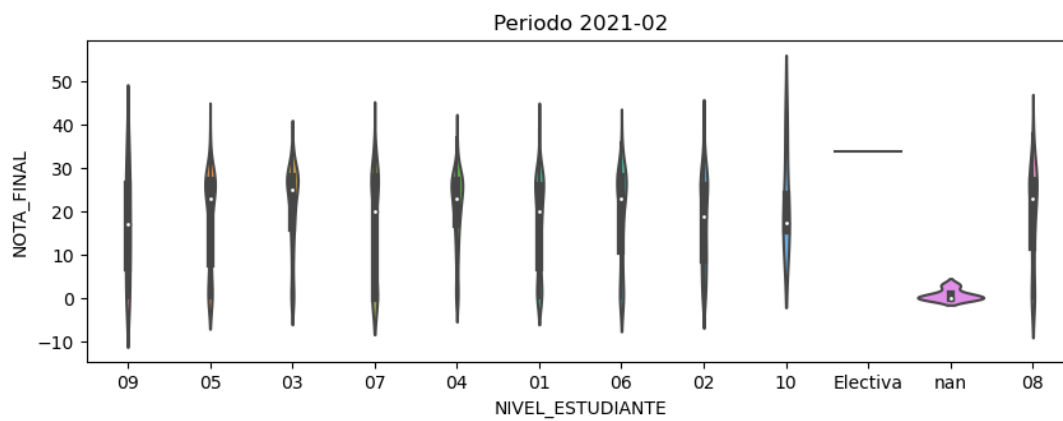


Ilustración 258

Diagrama de Violín del tercer periodo de la relación entre NIVEL_ESTUDIANTE y NOTA_FINAL del reporte de Asignaturas Perdidas

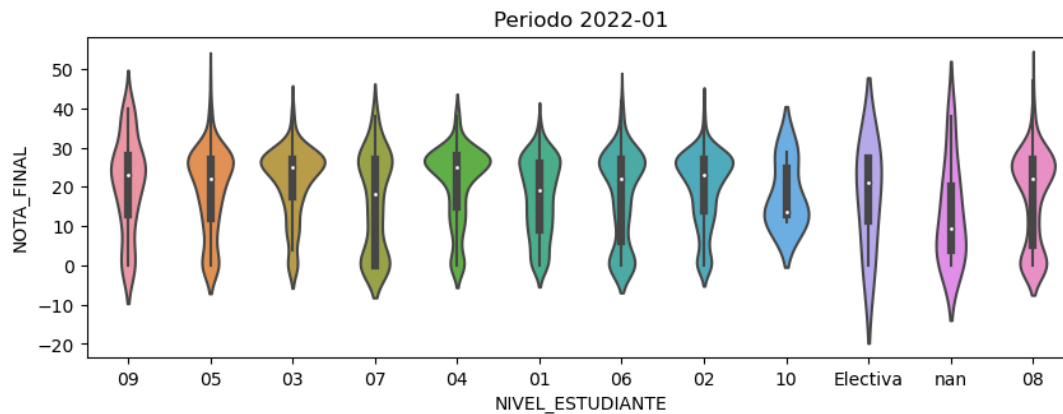
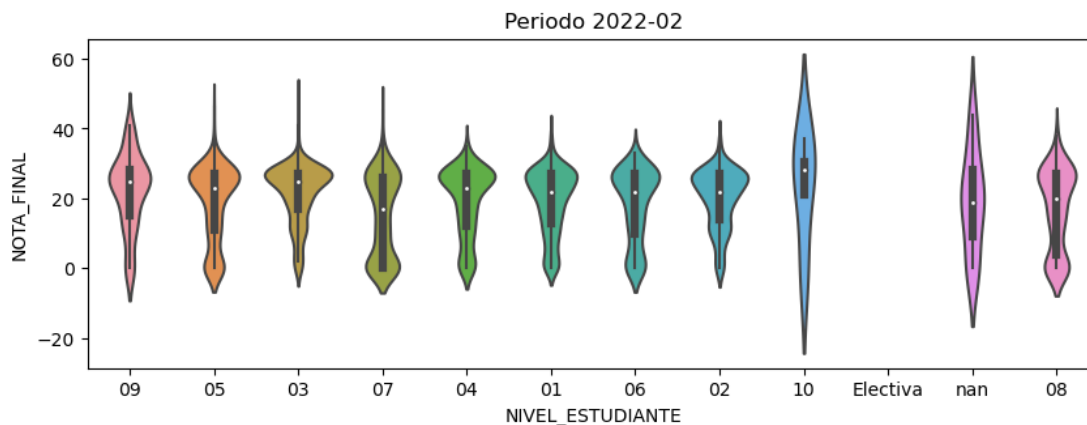


Ilustración 259

Diagrama de Violín del cuarto periodo de la relación entre NIVEL_ESTUDIANTE y NOTA_FINAL del reporte de Asignaturas Perdidas



Se puede concluir que es probable que exista una relación entre el nivel académico de un estudiante y su rendimiento al momento de reprobar, puesto que, a pesar de que varía la distribución entre periodos, el patrón de figura mantiene cierta semejanza por nivel; además, se puede observar cómo este patrón de figura de las distribuciones varía entre niveles, aunque en algunas ocasiones de manera muy discreta. Esto quiere decir que es bastante probable que exista una relación entre los niveles y el rendimiento académico.

Por otro lado, la relación entre las áreas de las asignaturas que cursan los estudiantes y sus notas finales será acatada mediante el uso de un diagrama de violín:

Ilustración 260

Código que genera los gráficos para la relación entre MATE y NOTA_FINAL del reporte de Asignaturas Perdidas

```
# Ordenamos Las categorías alfabéticamente
orden_categorias = set()
for perdida in perdidas:
    orden_categorias.update(list(sorted(perdida["MATE"].unique())))

# Grid para Los gráficos
fig, axes = plt.subplots(4, 1, figsize=(10, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='MATE', y='NOTA_FINAL', data=perdidas[0], order=orden_categorias, ax=axes[0])

axes[1].set_title('Periodo 2021-02')
sns.violinplot(x='MATE', y='NOTA_FINAL', data=perdidas[1], order=orden_categorias, ax=axes[1])

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='MATE', y='NOTA_FINAL', data=perdidas[2], order=orden_categorias, ax=axes[2])

axes[3].set_title('Periodo 2022-02')
sns.violinplot(x='MATE', y='NOTA_FINAL', data=perdidas[3], order=orden_categorias, ax=axes[3])

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.4)

plt.show() # Mostrar Los gráficos
```

Ilustración 261

Diagrama de Violín del primer periodo de la relación entre MATE y NOTA_FINAL del reporte de Asignaturas Perdidas

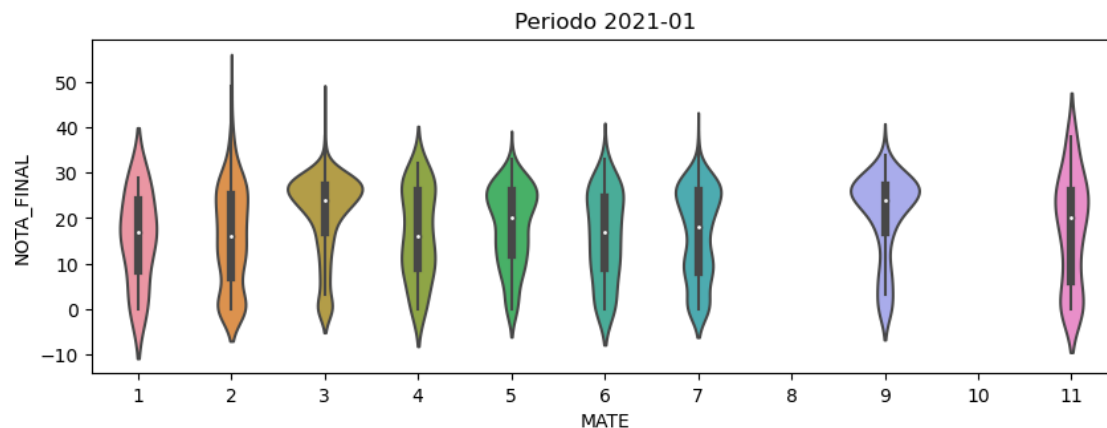


Ilustración 262

Diagrama de Violín del segundo periodo de la relación entre MATE y NOTA_FINAL del reporte de Asignaturas Perdidas

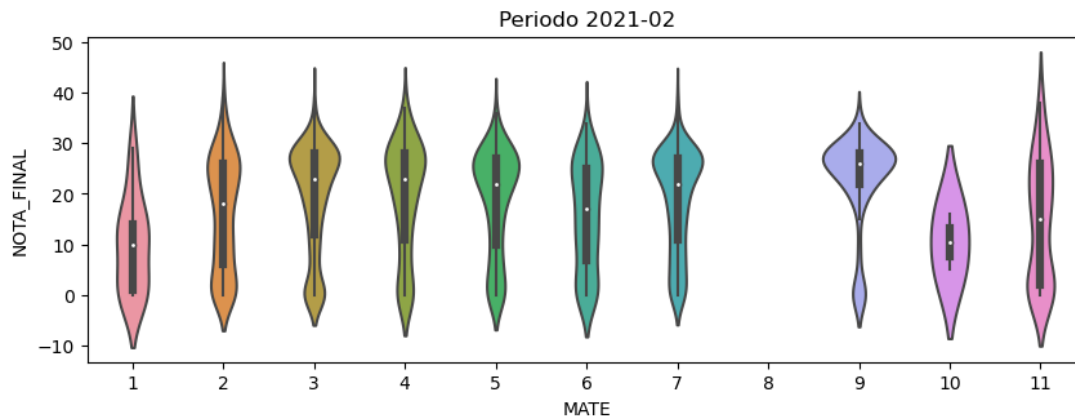


Ilustración 263

Diagrama de Violín del tercer periodo de la relación entre MATE y NOTA_FINAL del reporte de Asignaturas Perdidas

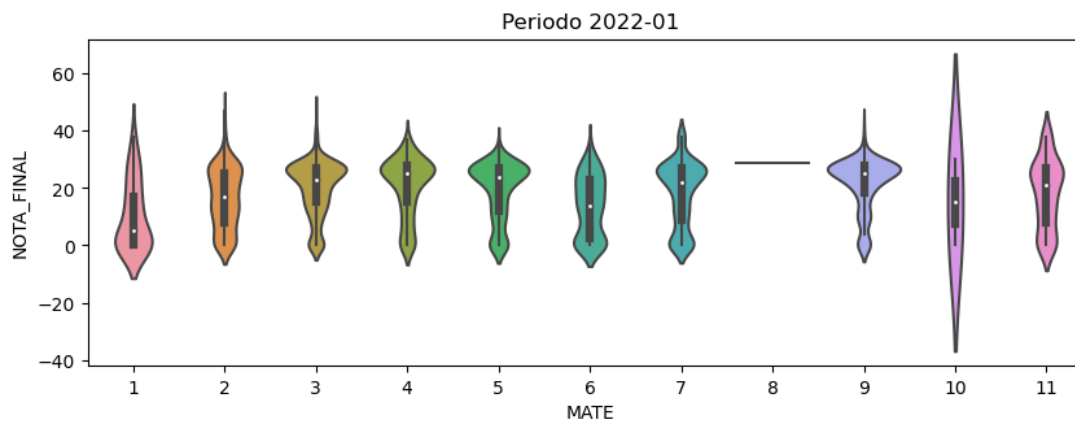
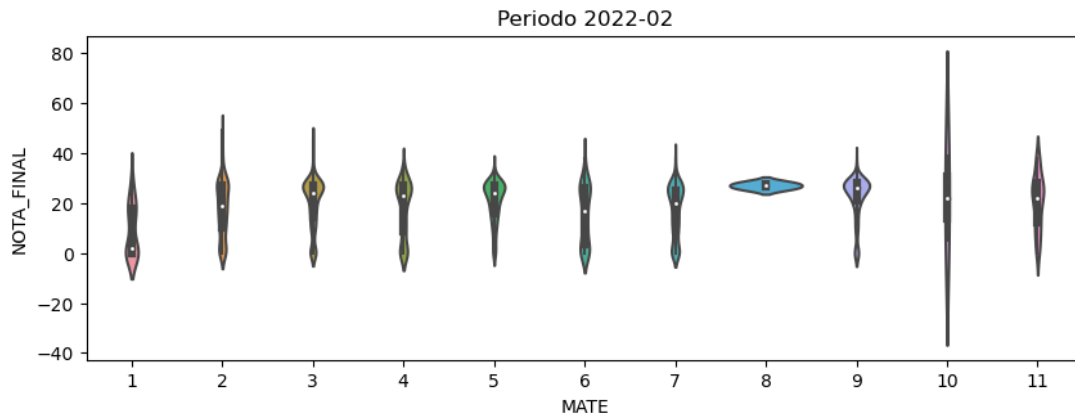


Ilustración 264

Diagrama de Violín del cuarto periodo de la relación entre MATE y NOTA_FINAL del reporte de Asignaturas Perdidas



En conclusión, es evidente una estrecha relación entre el área de asignatura y el desempeño del estudiante en ella, se puede observar como el patrón de figura se mantiene a través de los periodos académicos y, a su vez, son diferentes por cada área.

En sentido diferente, la exploración de la relación entre la cantidad de veces que un estudiante ha tomado un curso y su nota final se la realizará mediante el uso de un diagrama de violín:

Ilustración 265

Código que genera los gráficos para la relación entre VECES y NOTA_FINAL del reporte de Asignaturas Perdidas

```
# Ordenamos las categorías alfabéticamente
orden_categorias = set()
for perdida in perdidas:
    orden_categorias.update(list(sorted(perdida["VECES"].unique())))

# Grid para los gráficos
fig, axes = plt.subplots(4, 1, figsize=(10, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='VECES', y='NOTA_FINAL', data=perdidas[0], order=orden_categorias, ax=axes[0])

axes[1].set_title('Periodo 2021-02')
sns.violinplot(x='VECES', y='NOTA_FINAL', data=perdidas[1], order=orden_categorias, ax=axes[1])

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='VECES', y='NOTA_FINAL', data=perdidas[2], order=orden_categorias, ax=axes[2])

axes[3].set_title('Periodo 2022-02')
sns.violinplot(x='VECES', y='NOTA_FINAL', data=perdidas[3], order=orden_categorias, ax=axes[3])

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.4)

plt.show() # Mostrar los gráficos
```

Ilustración 266

Diagrama de Violín del primer periodo de la relación entre VECES y NOTA_FINAL del reporte de Asignaturas Perdidas

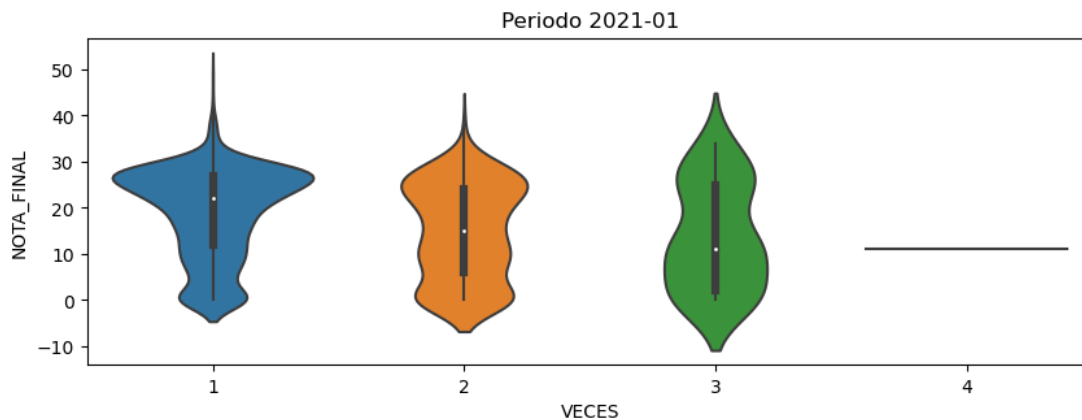


Ilustración 267

Diagrama de Violín del segundo periodo de la relación entre VECES y NOTA_FINAL del reporte de Asignaturas Perdidas

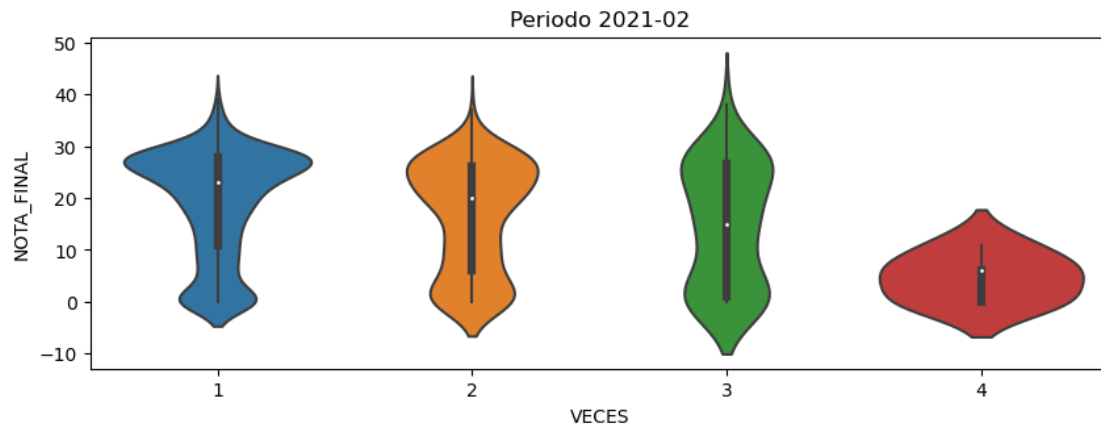


Ilustración 268

Diagrama de Violín del tercer periodo de la relación entre VECES y NOTA_FINAL del reporte de Asignaturas Perdidas

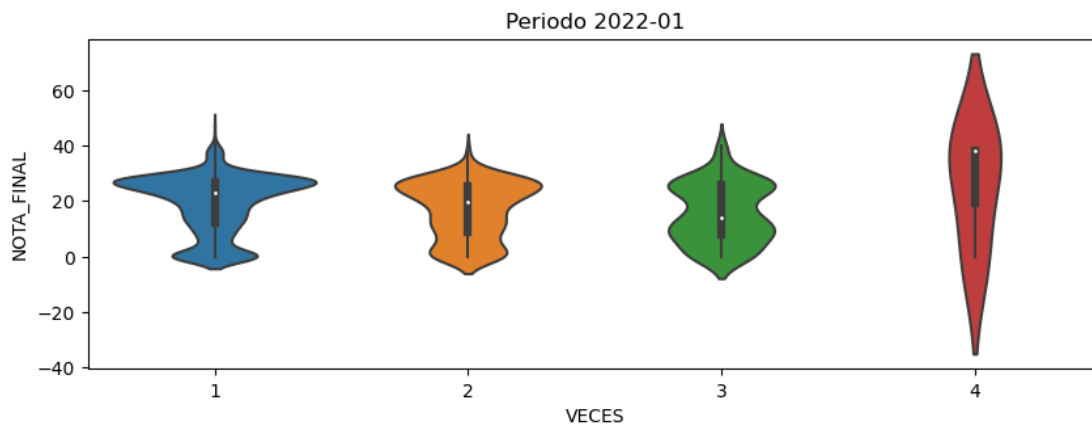
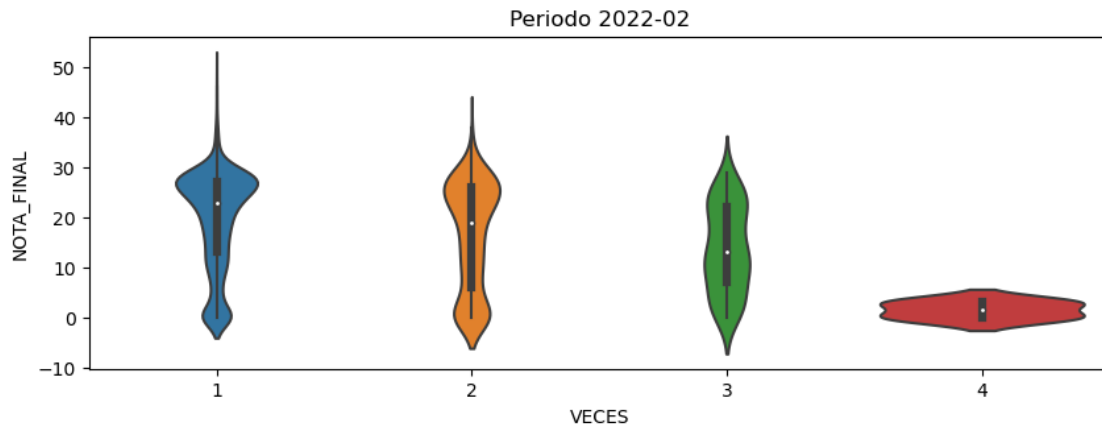


Ilustración 269

Diagrama de Violín del cuarto periodo de la relación entre VECES y NOTA_FINAL del reporte de Asignaturas Perdidas



En conclusión, se puede observar una prácticamente imperceptible relación entre la cantidad de veces en las que un estudiante repite una asignatura y su rendimiento académico, por lo que, posiblemente esta variable no tenga efecto en el rendimiento académico si la misma no presenta relevancia en los análisis de relación enfocados en ella.

En último término con lo que, respecto a la búsqueda de relaciones en función de las calificaciones finales de los estudiantes, es momento de presentar la relación entre los créditos de una asignatura que un estudiante está cursando y sus notas finales haciendo uso de un diagrama de caja y bigotes:

Ilustración 270

Código que genera los gráficos para la relación entre CRED y NOTA_FINAL del reporte de Asignaturas Perdidas

```
# Ordenamos las categorías alfabéticamente
orden_categorias = set()
for perdida in perdidas:
    orden_categorias.update(list(sorted(perdida["CRED"].unique())))

# Grid para los gráficos
fig, axes = plt.subplots(4, 1, figsize=(10, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.boxplot(x='CRED', y='NOTA_FINAL', data=perdidas[0], order=orden_categorias, ax=axes[0])

axes[1].set_title('Periodo 2021-02')
sns.boxplot(x='CRED', y='NOTA_FINAL', data=perdidas[1], order=orden_categorias, ax=axes[1])

axes[2].set_title('Periodo 2022-01')
sns.boxplot(x='CRED', y='NOTA_FINAL', data=perdidas[2], order=orden_categorias, ax=axes[2])

axes[3].set_title('Periodo 2022-02')
sns.boxplot(x='CRED', y='NOTA_FINAL', data=perdidas[3], order=orden_categorias, ax=axes[3])

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.4)

plt.show() # Mostrar los gráficos
```

Ilustración 271

Diagrama de Violín del primer periodo de la relación entre CRED y NOTA_FINAL del reporte de Asignaturas Perdidas

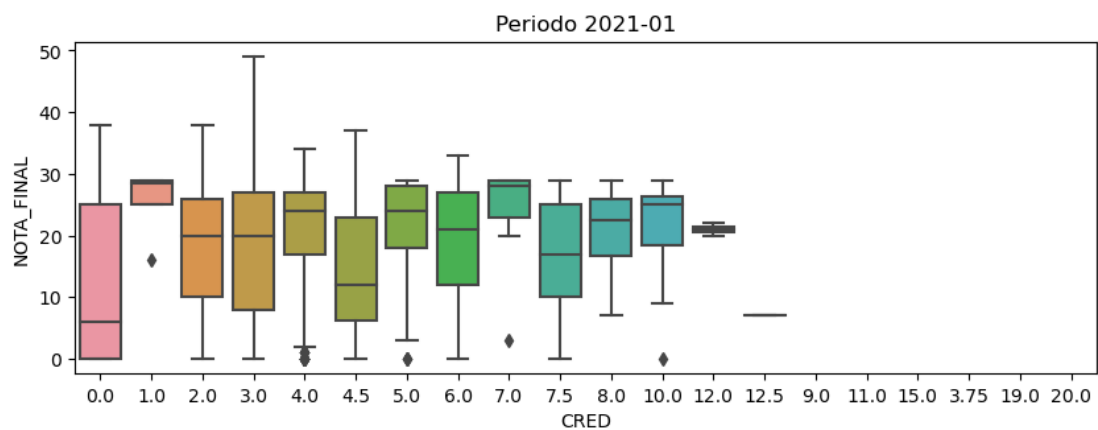


Ilustración 272

Diagrama de Violín del segundo periodo de la relación entre CRED y NOTA_FINAL del reporte de Asignaturas Perdidas

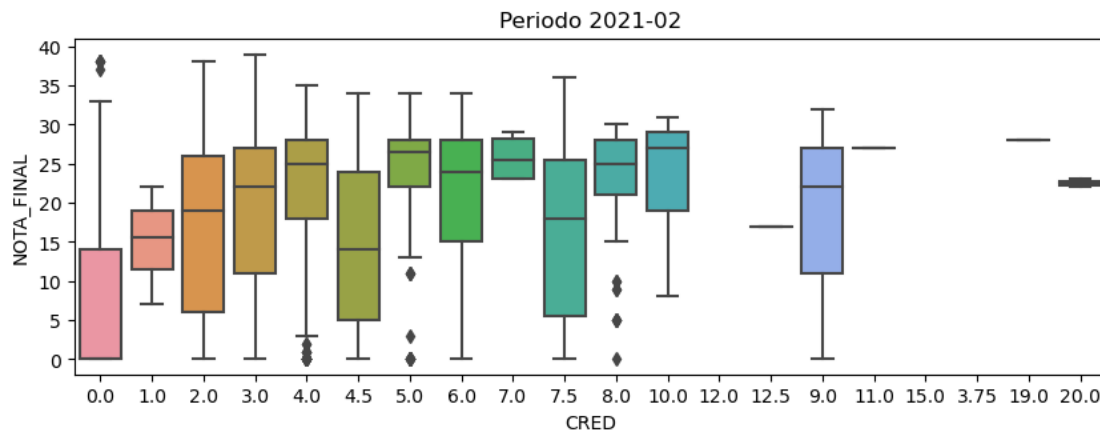


Ilustración 273

Diagrama de Violín del tercer periodo de la relación entre CRED y NOTA_FINAL del reporte de Asignaturas Perdidas

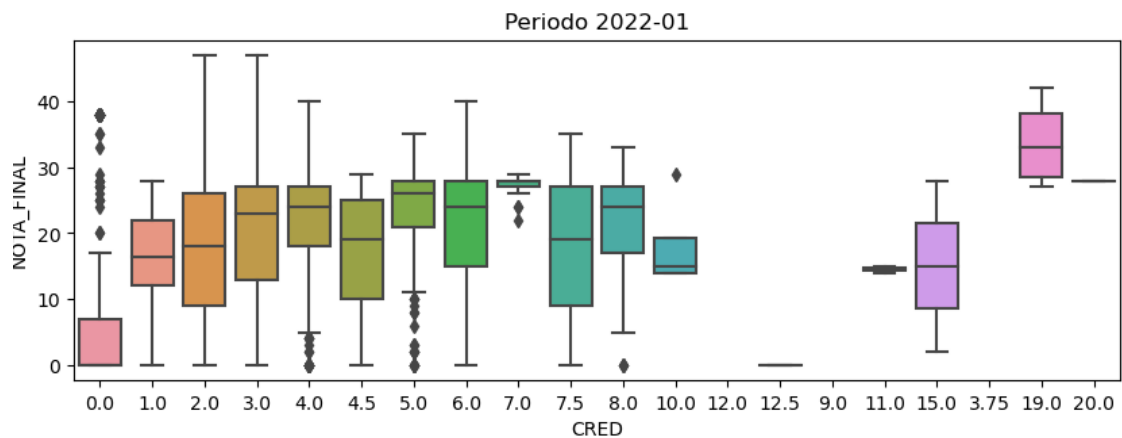
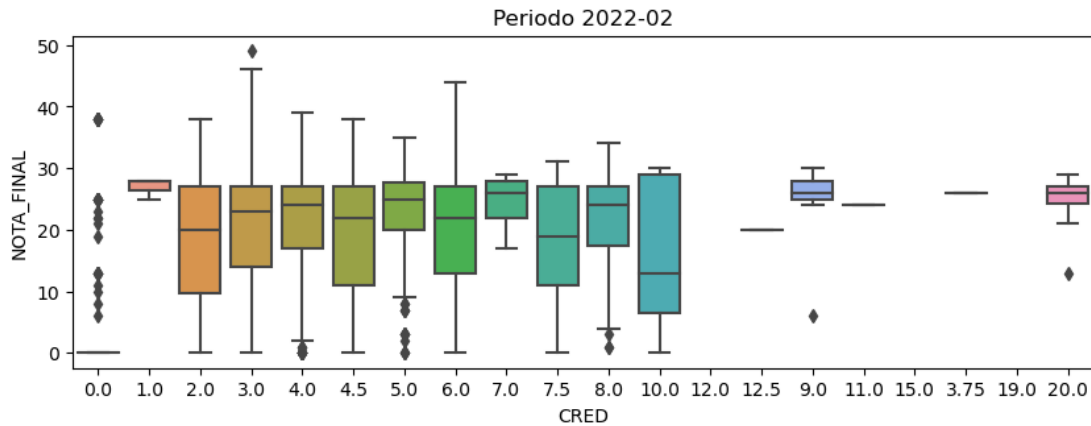


Ilustración 274

Diagrama de Violín del cuarto periodo de la relación entre CRED y NOTA_FINAL del reporte de Asignaturas Perdidas



En conclusión, se puede observar que no es posible la negación de la existencia de una relación entre el rendimiento académico y la cantidad de créditos de una asignatura, sin embargo, también es factible mencionar que a partir de las asignaturas con un peso superior a 10 créditos no presentan una relación. Debido a esto, es necesario evaluar la relación con un mapa de calor haciendo uso de una matriz de correlación.

En consecuencia, de haber terminado la evaluación de la primera hipótesis con respecto a las calificaciones finales, es momento de evaluar la hipótesis con respecto a la cantidad de veces que un estudiante repitió una misma asignatura. Para empezar, se evaluará entre los planes de estudios de un estudiante y la cantidad de veces que este repitió una asignatura utilizando un diagrama de violín:

Ilustración 275

Código que genera los gráficos para la relación entre PLAN_E y VECES del reporte de Asignaturas Perdidas

```
# Ordenamos las categorías alfabéticamente
orden_categorias = set()
for perdida in perdidas:
    orden_categorias.update(list(sorted(perdida["PLAN_E"].unique())))

# Grid para los gráficos
fig, axes = plt.subplots(4, 1, figsize=(10, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='PLAN_E', y='VECES', data=perdidas[0], order=orden_categorias, ax=axes[0])

axes[1].set_title('Periodo 2021-02')
sns.violinplot(x='PLAN_E', y='VECES', data=perdidas[1], order=orden_categorias, ax=axes[1])

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='PLAN_E', y='VECES', data=perdidas[2], order=orden_categorias, ax=axes[2])

axes[3].set_title('Periodo 2022-02')
sns.violinplot(x='PLAN_E', y='VECES', data=perdidas[3], order=orden_categorias, ax=axes[3])

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.4)

plt.show() # Mostrar los gráficos
```

Ilustración 276

Diagrama de Violín del primer periodo de la relación entre PLAN_E y VECES del reporte de Asignaturas Perdidas

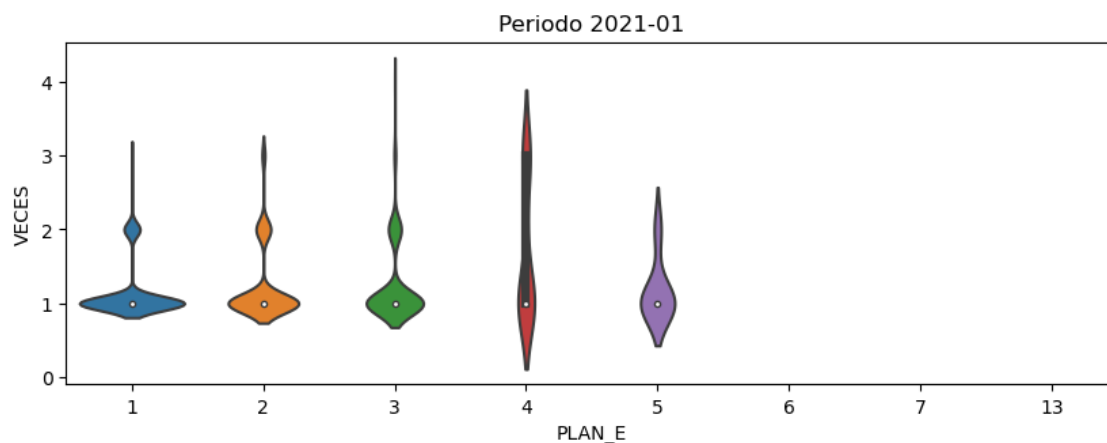


Ilustración 277

Diagrama de Violín del segundo periodo de la relación entre PLAN_E y VECES del reporte de Asignaturas Perdidas

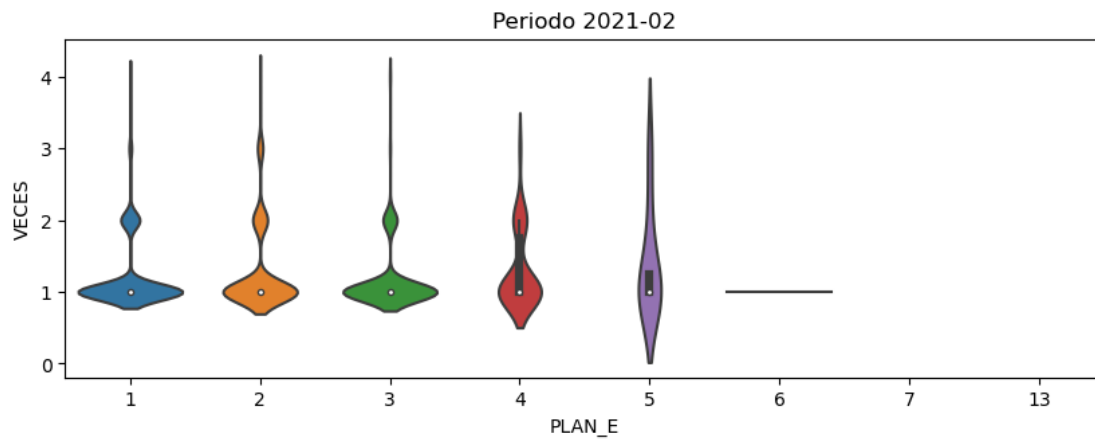


Ilustración 278

Diagrama de Violín del tercer periodo de la relación entre PLAN_E y VECES del reporte de Asignaturas Perdidas

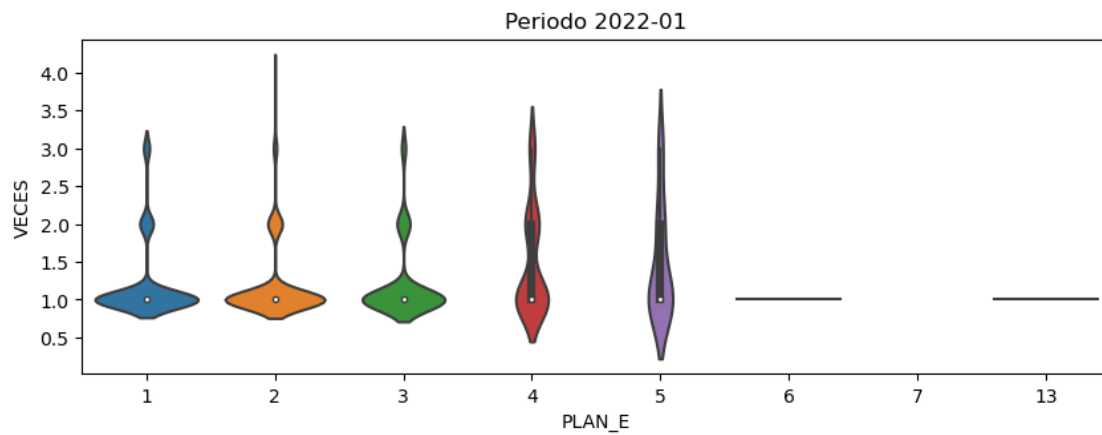
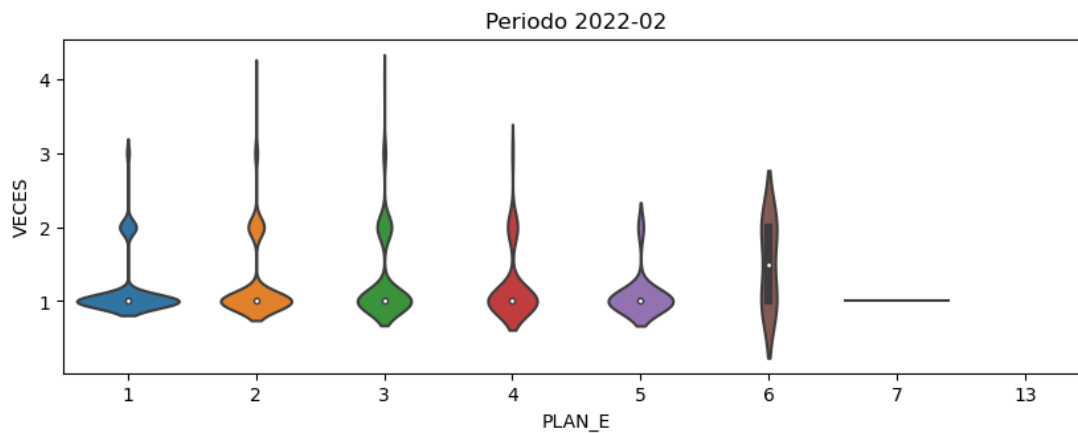


Ilustración 279

Diagrama de Violín del cuarto periodo de la relación entre PLAN_E y VECES del reporte de Asignaturas Perdidas



En conclusión, se puede observar que realmente no existe una relación aparente entre el plan de estudios y el rendimiento académico; no obstante, si bien, en algunos periodos podría argumentarse lo contrario, debido a su gran semejanza en las distribuciones y al patrón de figura, se ha decidido prescindir de esta variable para el modelo.

Como segunda propuesta para la hipótesis con respecto a la variable VECES, la relación entre las facultades y la cantidad de veces que los estudiantes repitieron una asignatura se realizará mediante los diagramas de violín siguientes:

Ilustración 280

Código que genera los gráficos para la relación entre DESC_FACULTAD y VECES del reporte de Asignaturas Perdidas

```
# Ordenamos las categorías alfabéticamente
orden_categorias = set()
for perdida in perdidas:
    orden_categorias.update(list(sorted(perdida["DESC_FACULTAD"].unique())))

# Grid para los gráficos
fig, axes = plt.subplots(4, 1, figsize=(10, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='DESC_FACULTAD', y='VECES', data=perdidas[0], order=orden_categorias, ax=axes[0])
axes[0].set_xticklabels(axes[0].get_xticklabels(), rotation=90)

axes[1].set_title('Periodo 2021-02')
sns.violinplot(x='DESC_FACULTAD', y='VECES', data=perdidas[1], order=orden_categorias, ax=axes[1])
axes[1].set_xticklabels(axes[1].get_xticklabels(), rotation=90)

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='DESC_FACULTAD', y='VECES', data=perdidas[2], order=orden_categorias, ax=axes[2])
axes[2].set_xticklabels(axes[2].get_xticklabels(), rotation=90)

axes[3].set_title('Periodo 2022-02')
sns.violinplot(x='DESC_FACULTAD', y='VECES', data=perdidas[3], order=orden_categorias, ax=axes[3])
axes[3].set_xticklabels(axes[3].get_xticklabels(), rotation=90)

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=2.4)

plt.show() # Mostrar los gráficos
```

Ilustración 281

Diagrama de Violín del primer periodo de la relación entre DESC_FACULTAD y VECES del reporte de Asignaturas Perdidas

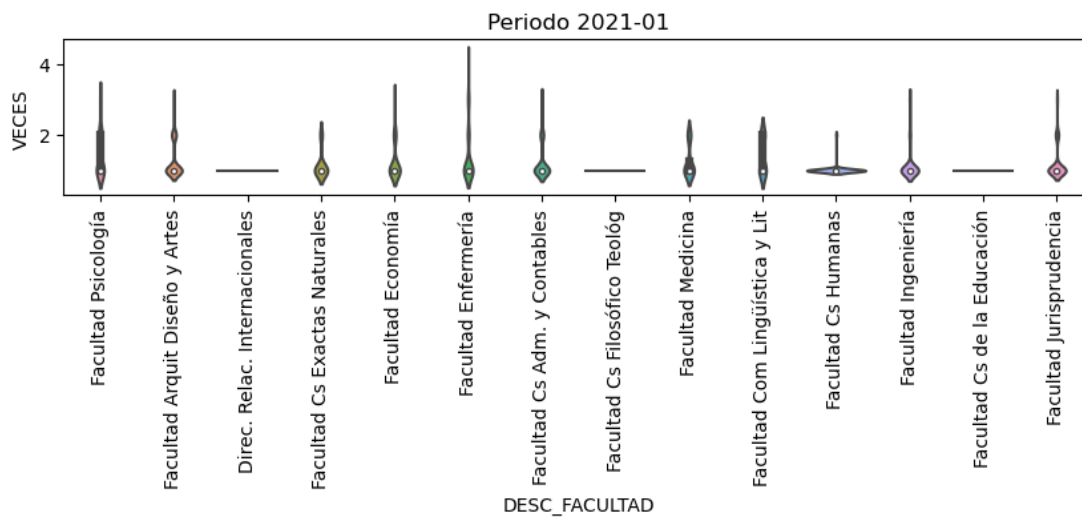


Ilustración 282

Diagrama de Violín del segundo periodo de la relación entre DESC_FACULTAD y VECES del reporte de Asignaturas Perdidas

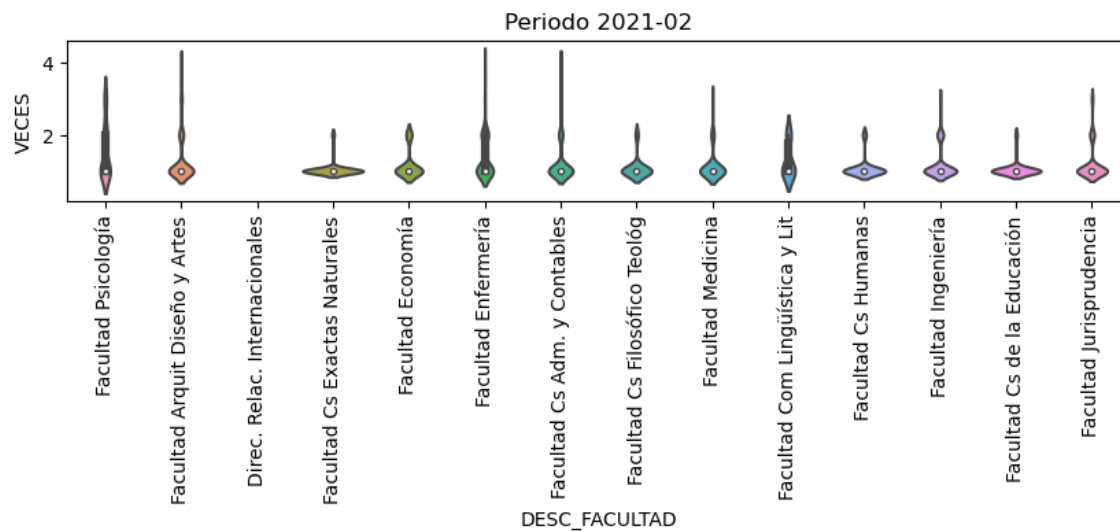


Ilustración 283

Diagrama de Violín del tercer periodo de la relación entre DESC_FACULTAD y VECES del reporte de Asignaturas Perdidas

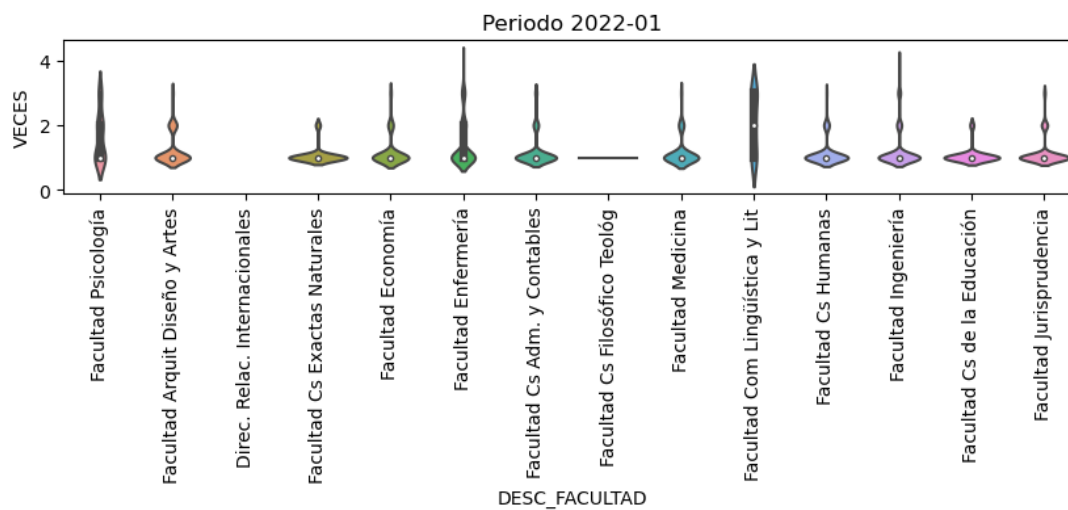
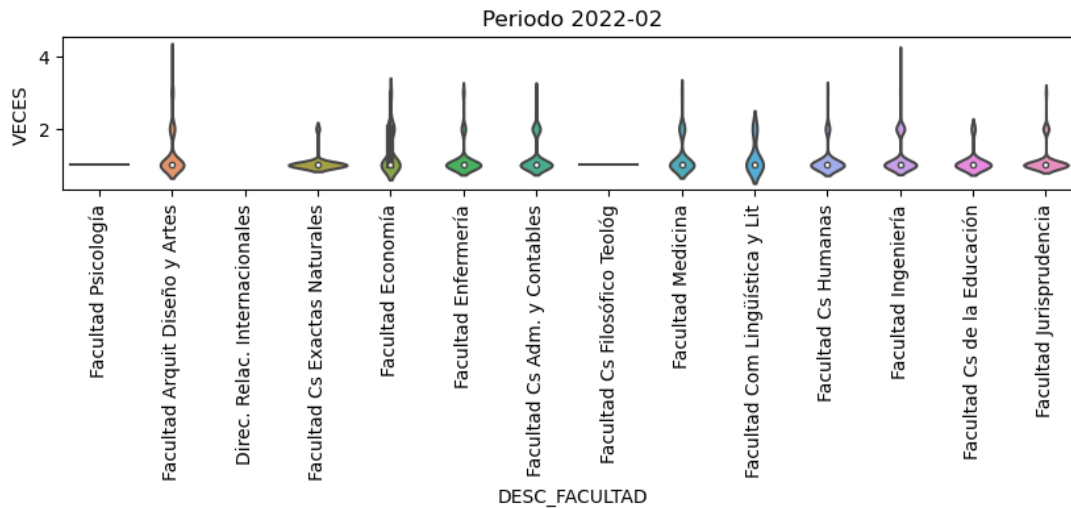


Ilustración 284

Diagrama de Violín del cuarto periodo de la relación entre DESC_FACULTAD y VECES del reporte de Asignaturas Perdidas



En conclusión, se puede observar de igual manera una estrecha relación entre las facultades y el rendimiento académico, en esta ocasión con respecto a la cantidad de veces que un estudiante reprobado repite una misma asignatura, es decir, es concreta la existencia una relación entre las facultades y el rendimiento académico universitario.

Como exploración siguiente, la relación entre las carreras que pertenecen los estudiantes y la cantidad de veces que repiten una misma asignatura se realizará mediante los consiguientes diagramas de caja y bigotes:

Ilustración 285

Código que genera los gráficos para la relación entre DESC_PROGRAMA y VECES del reporte de Asignaturas Perdidas

```
# Ordenamos las categorías alfabéticamente
orden_categorias = set()
for perdida in perdidas:
    orden_categorias.update(list(sorted(perdida["DESC_PROGRAMA"].unique())))

# Grid para los gráficos
fig, axes = plt.subplots(4, 1, figsize=(12, 25))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.boxplot(x='DESC_PROGRAMA', y='VECES', data=perdidas[0], order=orden_categorias, ax=axes[0])
axes[0].set_xticklabels(axes[0].get_xticklabels(), rotation=90)

axes[1].set_title('Periodo 2021-02')
sns.boxplot(x='DESC_PROGRAMA', y='VECES', data=perdidas[1], order=orden_categorias, ax=axes[1])
axes[1].set_xticklabels(axes[1].get_xticklabels(), rotation=90)

axes[2].set_title('Periodo 2022-01')
sns.boxplot(x='DESC_PROGRAMA', y='VECES', data=perdidas[2], order=orden_categorias, ax=axes[2])
axes[2].set_xticklabels(axes[2].get_xticklabels(), rotation=90)

axes[3].set_title('Periodo 2022-02')
sns.boxplot(x='DESC_PROGRAMA', y='VECES', data=perdidas[3], order=orden_categorias, ax=axes[3])
axes[3].set_xticklabels(axes[3].get_xticklabels(), rotation=90)

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=1.4)

plt.show() # Mostrar los gráficos
```

Ilustración 286

Diagrama de Violín del primer periodo de la relación entre DESC_PROGRAMA y VECES del reporte de Asignaturas Perdidas

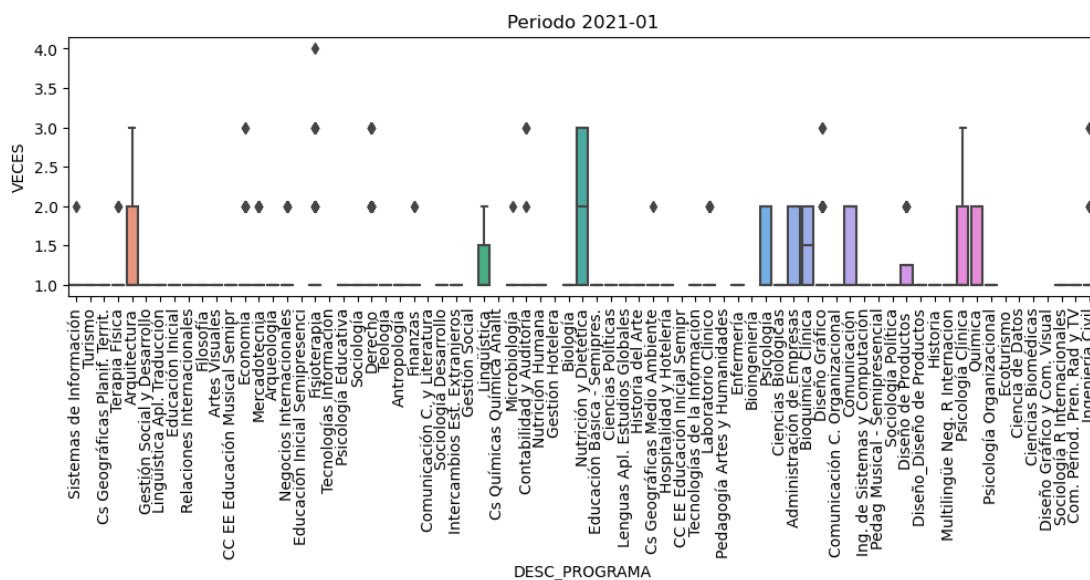


Ilustración 287

Diagrama de Violín del segundo periodo de la relación entre DESC_PROGRAMA y VECES del reporte de Asignaturas Perdidas

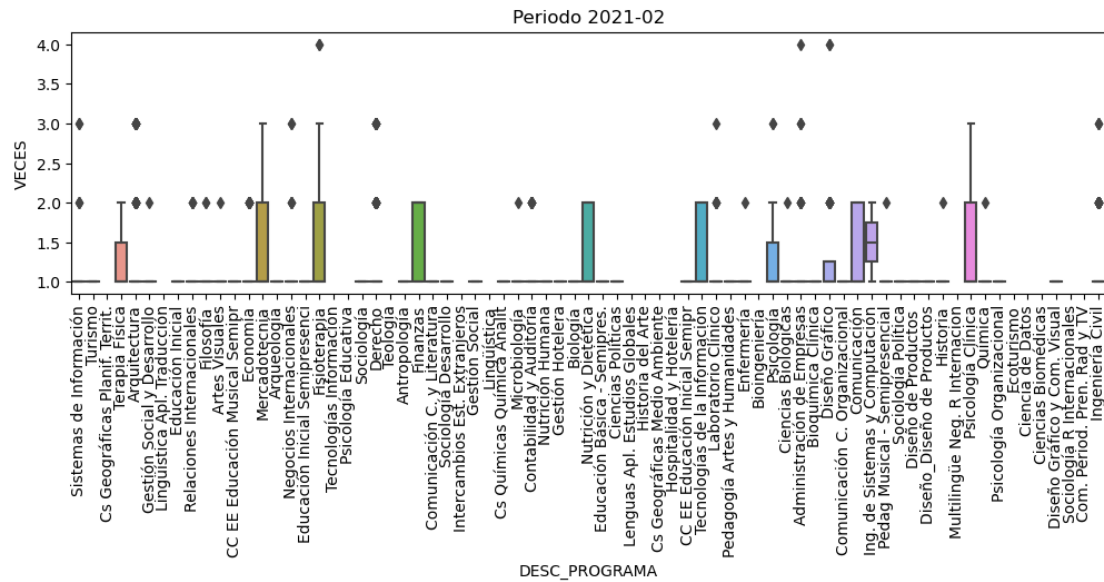


Ilustración 288

Diagrama de Violín del tercer periodo de la relación entre DESC_PROGRAMA y VECES del reporte de Asignaturas Perdidas

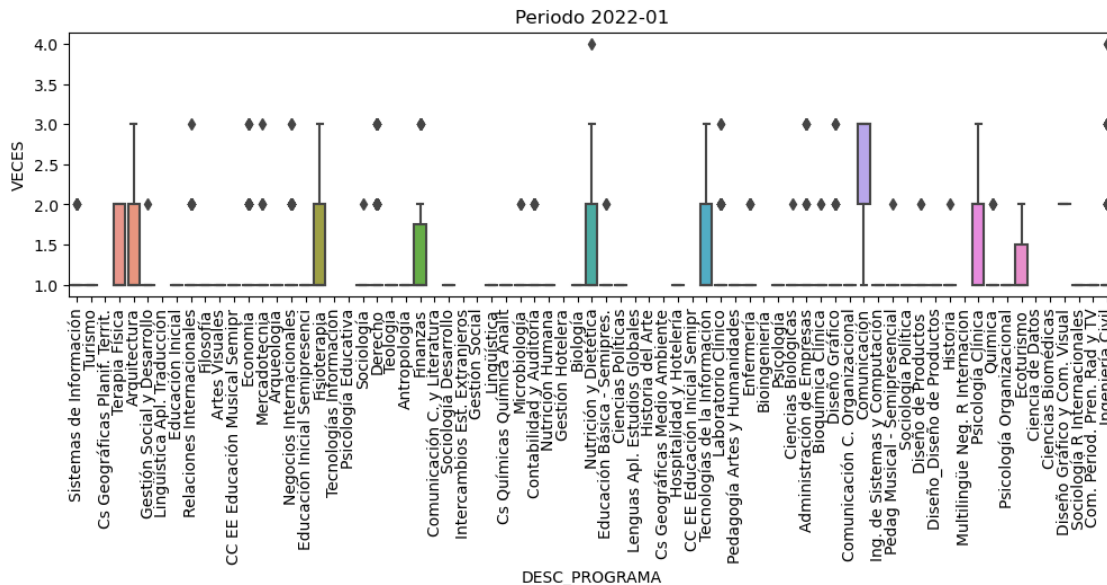
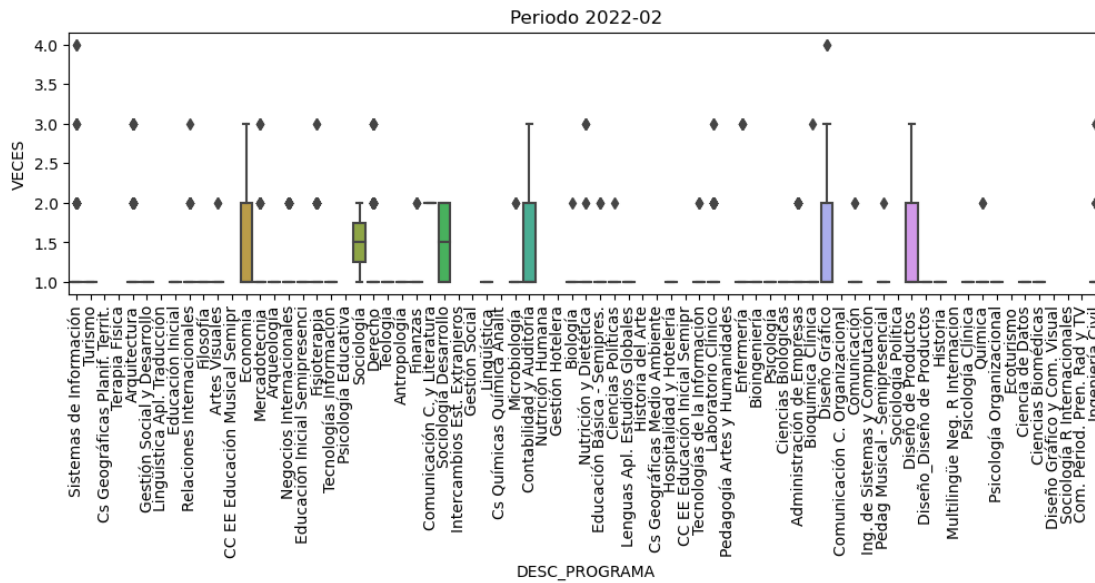


Ilustración 289

Diagrama de Violín del cuarto periodo de la relación entre DESC_PROGRAMA y VECES del reporte de Asignaturas Perdidas



En conclusión, a pesar de observar grandes distribuciones de datos en algunas carreras dado el periodo, no existe una aparente relación entre la cantidad de veces que un estudiante repite una asignatura y la carrera que se encuentra cursando.

En sentido diferente, la relación entre los niveles de los estudiantes y la cantidad de veces que han repetido una misma asignatura se podrá observar a continuación mediante diagramas de violín:

Ilustración 290

Código que genera los gráficos para la relación entre NIVEL_ESTUDIANTE y VECES del reporte de Asignaturas Perdidas

```
# Ordenamos Las categorías alfabéticamente
orden_categorias = set()
for perdida in perdidas:
    orden_categorias.update(list(sorted(perdida["NIVEL_ESTUDIANTE"].unique()))))

# Grid para Los gráficos
fig, axes = plt.subplots(4, 1, figsize=(10, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='NIVEL_ESTUDIANTE', y='VECES', data=perdidas[0], order=orden_categorias, ax=axes[0])

axes[1].set_title('Periodo 2021-02')
sns.violinplot(x='NIVEL_ESTUDIANTE', y='VECES', data=perdidas[1], order=orden_categorias, ax=axes[1])

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='NIVEL_ESTUDIANTE', y='VECES', data=perdidas[2], order=orden_categorias, ax=axes[2])

axes[3].set_title('Periodo 2022-02')
sns.violinplot(x='NIVEL_ESTUDIANTE', y='VECES', data=perdidas[3], order=orden_categorias, ax=axes[3])

# Ajustar el espacio entre Los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.4)

plt.show() # Mostrar Los gráficos
```

Ilustración 291

Diagrama de Violín del primer periodo de la relación entre NIVEL_ESTUDIANTE y VECES del reporte de Asignaturas Perdidas

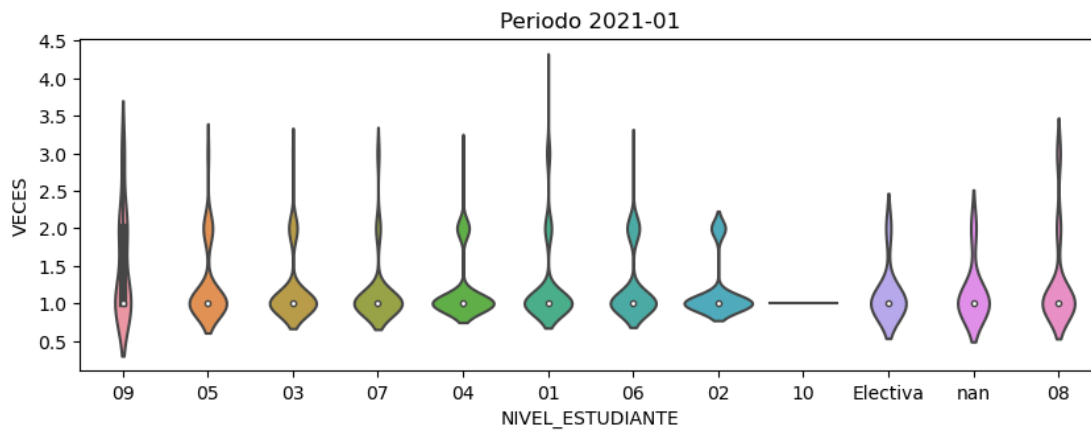


Ilustración 292

Diagrama de Violín del segundo periodo de la relación entre NIVEL_ESTUDIANTE y VECES del reporte de Asignaturas Perdidas

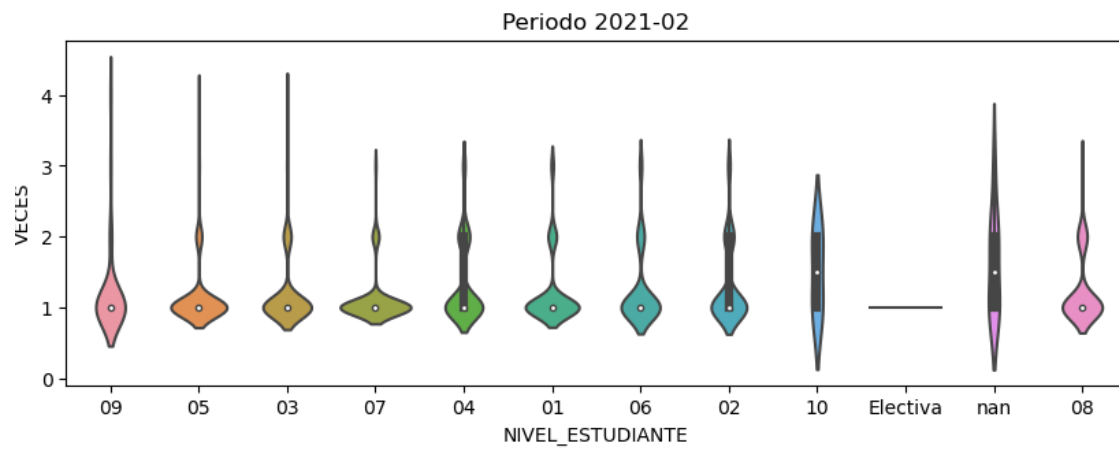


Ilustración 293

Diagrama de Violín del tercer periodo de la relación entre NIVEL_ESTUDIANTE y VECES del reporte de Asignaturas Perdidas

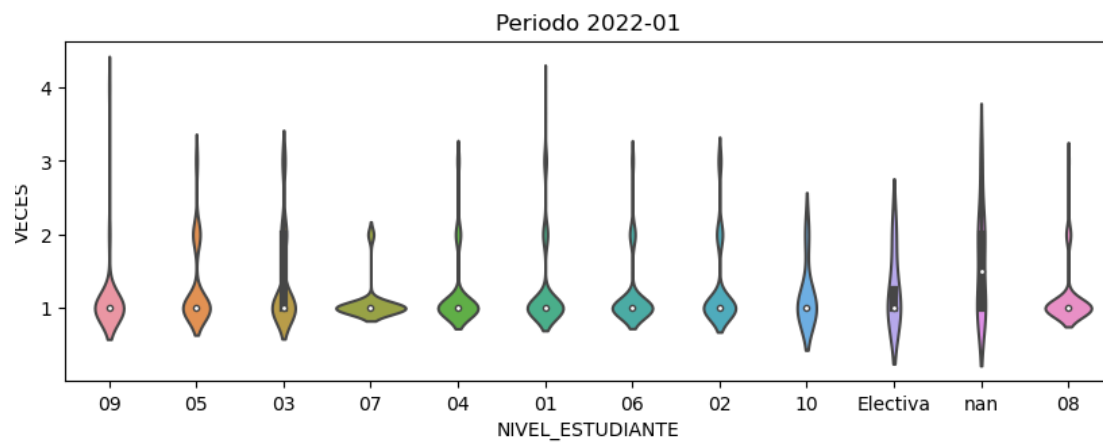
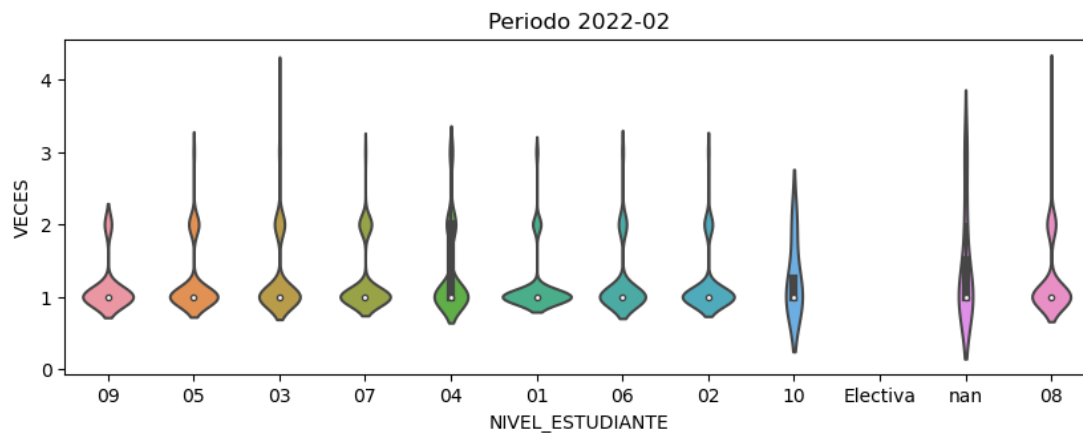


Ilustración 294

Diagrama de Violín del cuarto periodo de la relación entre NIVEL_ESTUDIANTE y VECES del reporte de Asignaturas Perdidas



En conclusión, se puede observar que tampoco existe una relación entre la cantidad de veces que un estudiante repite una asignatura y su nivel académico, ya que el patrón de figura y las distribuciones son prácticamente iguales entre los diferentes niveles académicos.

Por otra parte, considerando la relación entre las áreas de las asignaturas y la cantidad de veces que los estudiantes han repetido una asignatura de la misma área se podrá visualizar a continuación mediante el empleo de diagramas de violín:

Ilustración 295

Código que genera los gráficos para la relación entre MATE y VECES del reporte de Asignaturas Perdidas

```
# Ordenamos las categorías alfabéticamente
orden_categorias = set()
for perdida in perdidas:
    orden_categorias.update(list(sorted(perdida["MATE"].unique())))

# Grid para los gráficos
fig, axes = plt.subplots(4, 1, figsize=(10, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='MATE', y='VECES', data=perdidas[0], order=orden_categorias, ax=axes[0])

axes[1].set_title('Periodo 2021-02')
sns.violinplot(x='MATE', y='VECES', data=perdidas[1], order=orden_categorias, ax=axes[1])

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='MATE', y='VECES', data=perdidas[2], order=orden_categorias, ax=axes[2])

axes[3].set_title('Periodo 2022-02')
sns.violinplot(x='MATE', y='VECES', data=perdidas[3], order=orden_categorias, ax=axes[3])

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.4)

plt.show() # Mostrar los gráficos
```

Ilustración 296

Diagrama de Violín del primer periodo de la relación entre MATE y VECES del reporte de Asignaturas Perdidas

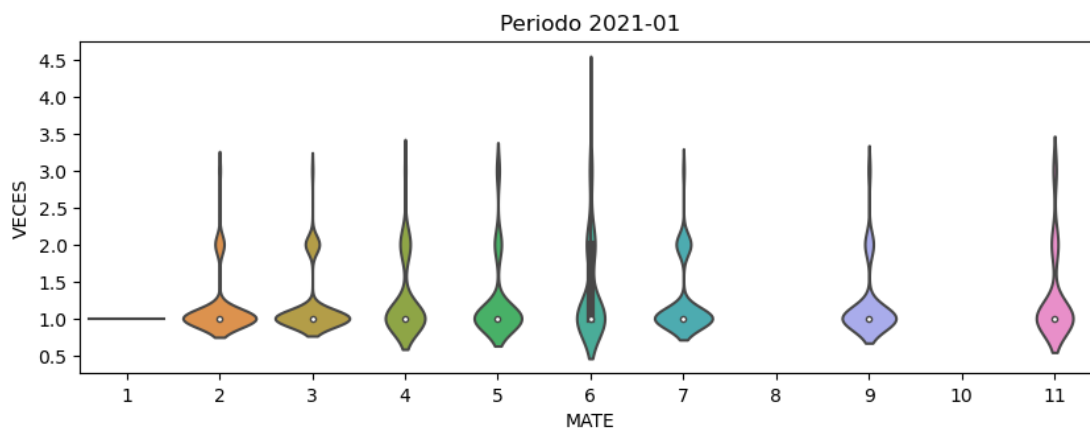


Ilustración 297

Diagrama de Violín del segundo periodo de la relación entre MATE y VECES del reporte de Asignaturas Perdidas

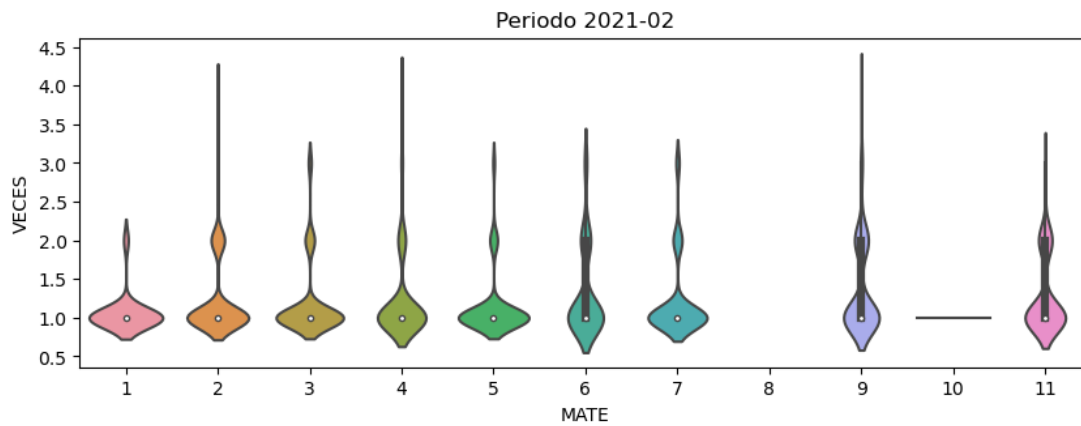


Ilustración 298

Diagrama de Violín del tercer periodo de la relación entre MATE y VECES del reporte de Asignaturas Perdidas

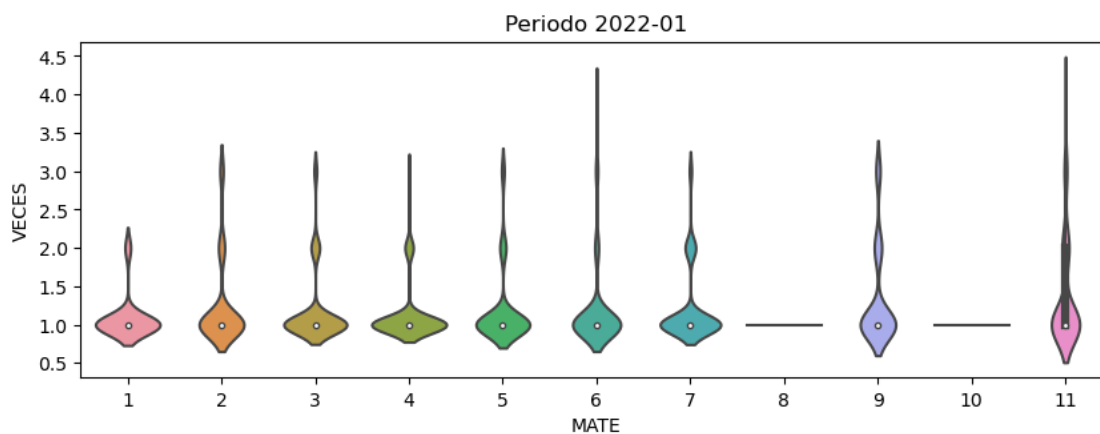
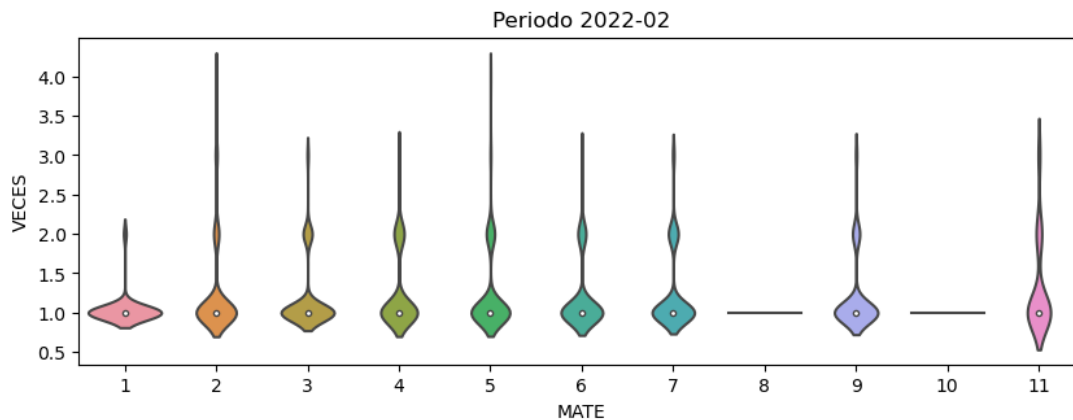


Ilustración 299

Diagrama de Violín del cuarto periodo de la relación entre MATE y VECES del reporte de Asignaturas Perdidas



En conclusión, se puede observar que no existe una estrecha relación entre las áreas de asignaturas y la cantidad de veces que se repite una asignatura en el área determinada, esto quiere decir que el pertenecer a un área de asignatura u otra, no refleja realmente si sus asignaturas correspondientes tendrán más o menos repeticiones, pues, es visible que en todas las áreas el comportamiento con respecto a repeticiones es igual.

Como último término para responder la hipótesis con respecto a relaciones con la cantidad de veces que un estudiante repite una misma asignatura, se considerará la evaluación entre los créditos de una asignatura y la cantidad de veces que un estudiante repitió dicha asignatura mediante diagramas de violín:

Ilustración 300

Código que genera los gráficos para la relación entre CRED y VECES del reporte de Asignaturas Perdidas

```
# Ordenamos las categorías alfabéticamente
orden_categorias = set()
for perdida in perdidas:
    orden_categorias.update(list(sorted(perdida["CRED"].unique())))

# Grid para los gráficos
fig, axes = plt.subplots(4, 1, figsize=(10, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='CRED', y='VECES', data=perdidas[0], order=orden_categorias, ax=axes[0])

axes[1].set_title('Periodo 2021-02')
sns.violinplot(x='CRED', y='VECES', data=perdidas[1], order=orden_categorias, ax=axes[1])

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='CRED', y='VECES', data=perdidas[2], order=orden_categorias, ax=axes[2])

axes[3].set_title('Periodo 2022-02')
sns.violinplot(x='CRED', y='VECES', data=perdidas[3], order=orden_categorias, ax=axes[3])

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.4)

plt.show() # Mostrar Los gráficos
```

Ilustración 301

Diagrama de Violín del primer periodo de la relación entre CRED y VECES del reporte de Asignaturas Perdidas

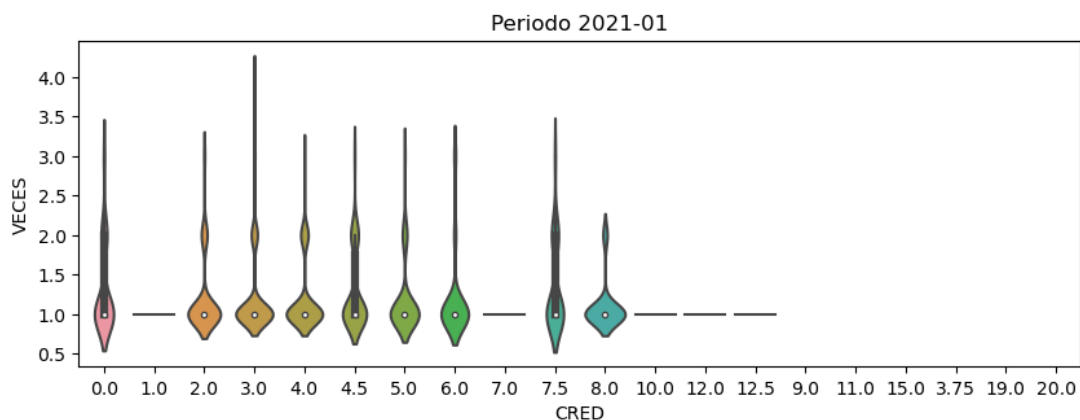


Ilustración 302

Diagrama de Violín del segundo periodo de la relación entre CRED y VECES del reporte de Asignaturas Perdidas

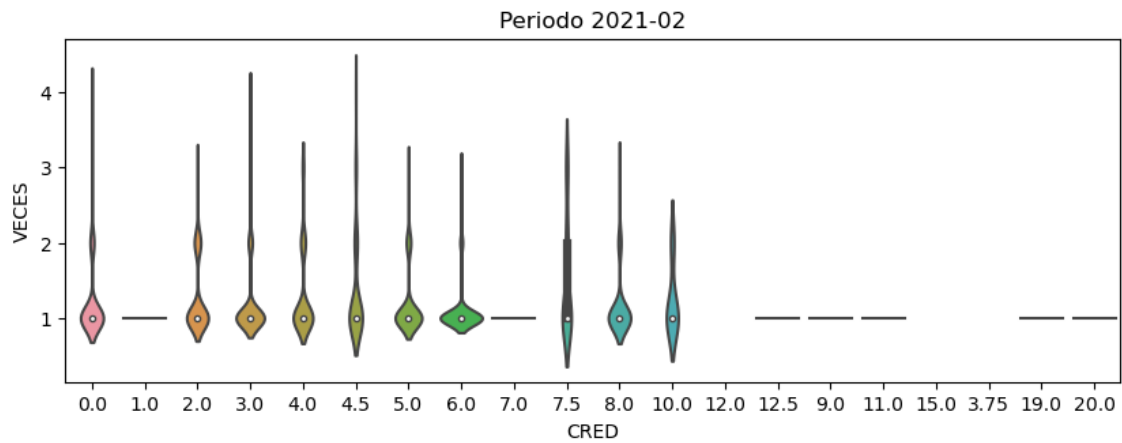


Ilustración 303

Diagrama de Violín del tercer periodo de la relación entre CRED y VECES del reporte de Asignaturas Perdidas

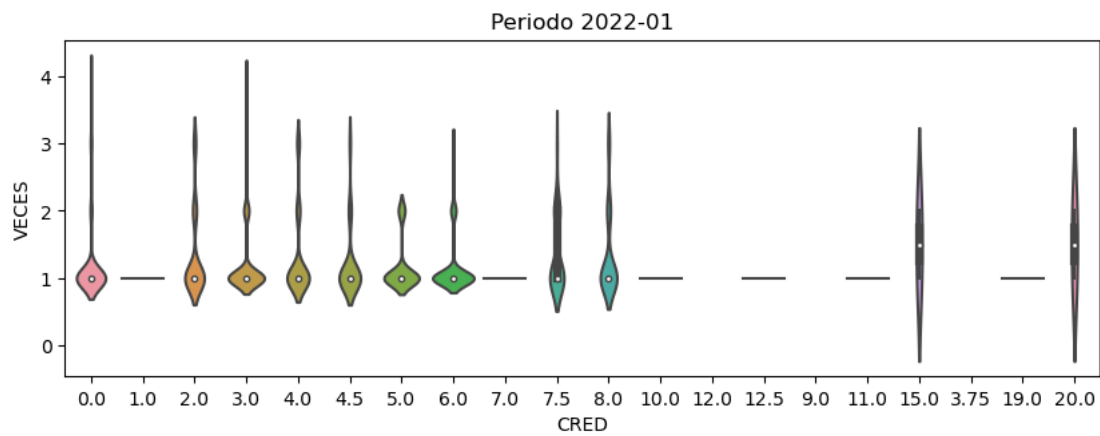
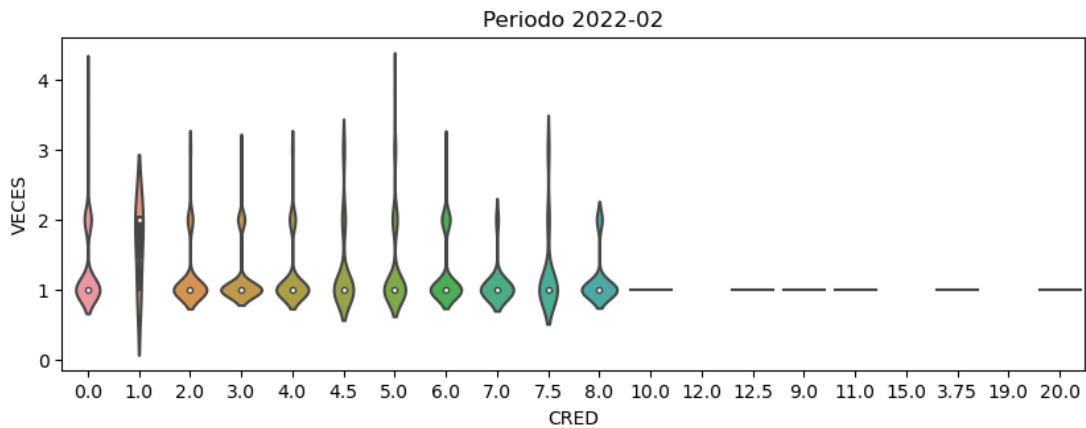


Ilustración 304

Diagrama de Violín del cuarto periodo de la relación entre CRED y VECES del reporte de Asignaturas Perdidas



En conclusión, se puede observar que no existe una relación clara entre la cantidad de veces que un estudiante repite una determinada asignatura y la cantidad de créditos que esta tiene.

En resumen, se han evaluado las hipótesis que sugieren las relaciones de diversos factores con respecto a las calificaciones finales y cantidad de veces que un estudiante ha cursado una misma asignatura, sin embargo, dado las conclusiones de cada análisis, se ve la necesidad de analizar más profundamente las variables de calificaciones finales, cantidad de veces y, también, la referente a créditos. Para esto, se emplearán gráficos de calor a base de matrices de correlación, se debe tener en cuenta que, en términos generales, los resultados más cercanos a 1 o -1 significan una correlación fuerte. El código, gráficos y conclusiones se puede encontrar a continuación:

Ilustración 305

Código que genera los gráficos para las variables NOTA_FINAL, CRED y VECES del reporte de Asignaturas Perdidas

```
# Grid para Los gráficos
fig, axes = plt.subplots(4, 1, figsize=(10, 17))

count = 0

periodos = ["2021-01", "2021-02", "2022-01", "2022-02"]

for perdida in perdidas:
    # Se crea un dataframe temporal para realizar cambios
    temp_df = perdida.copy()

    # Se convierten las columnas a tipos numéricos para evitar posibles errores
    temp_df['NOTA_FINAL'] = pd.to_numeric(temp_df['NOTA_FINAL'])
    temp_df['CRED'] = pd.to_numeric(temp_df['CRED'])
    temp_df['VECES'] = pd.to_numeric(temp_df['VECES'])

    # Se calcula la matriz de correlación
    corr_matrix = temp_df[['NOTA_FINAL', 'CRED', 'VECES']].corr()

    # Se crea la figura
    axes[count].set_title(f"Mapa de calor del periodo {periodos[count]}")
    sns.heatmap(corr_matrix, annot=True, cmap="coolwarm", ax=axes[count])
    count += 1

plt.tight_layout()
plt.show()
```

Ilustración 306

Mapa de calor del primer periodo de las variables NOTA_FINAL, CRED y VECES del reporte de Asignaturas Perdidas

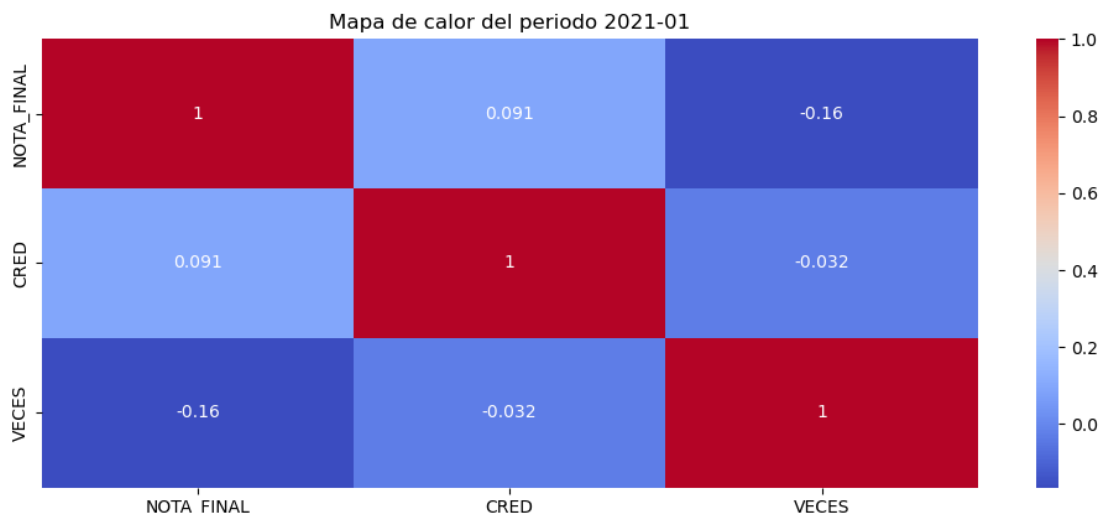


Ilustración 307

Mapa de calor del segundo periodo de las variables NOTA_FINAL, CRED y VECES del reporte de Asignaturas Perdidas

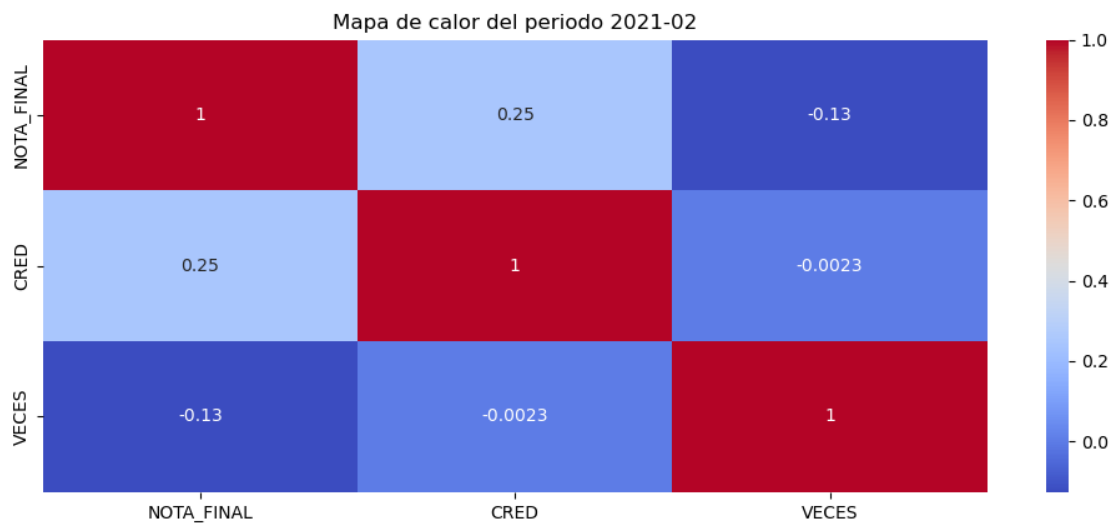


Ilustración 308

Mapa de calor del tercer periodo de las variables NOTA_FINAL, CRED y VECES del reporte de Asignaturas Perdidas

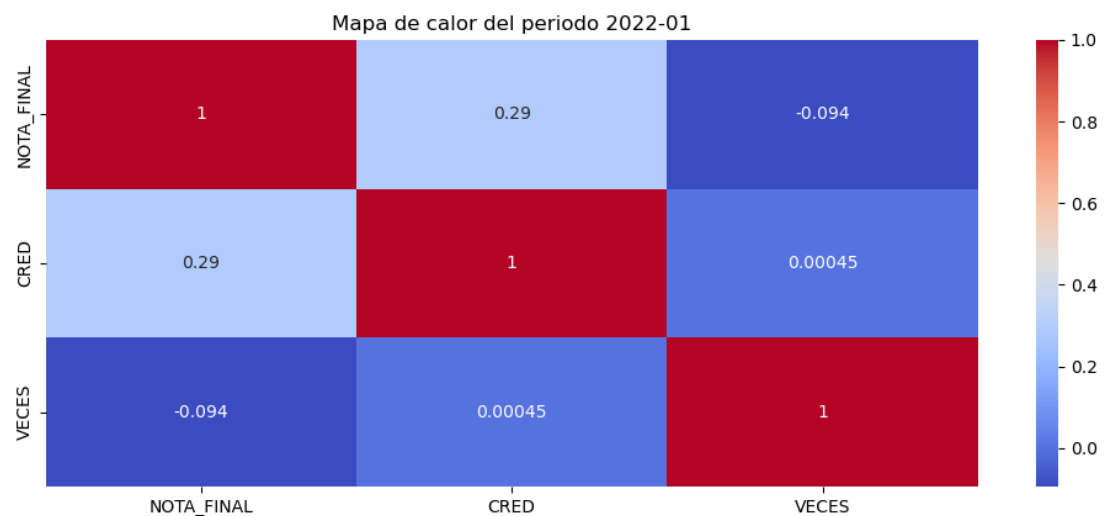
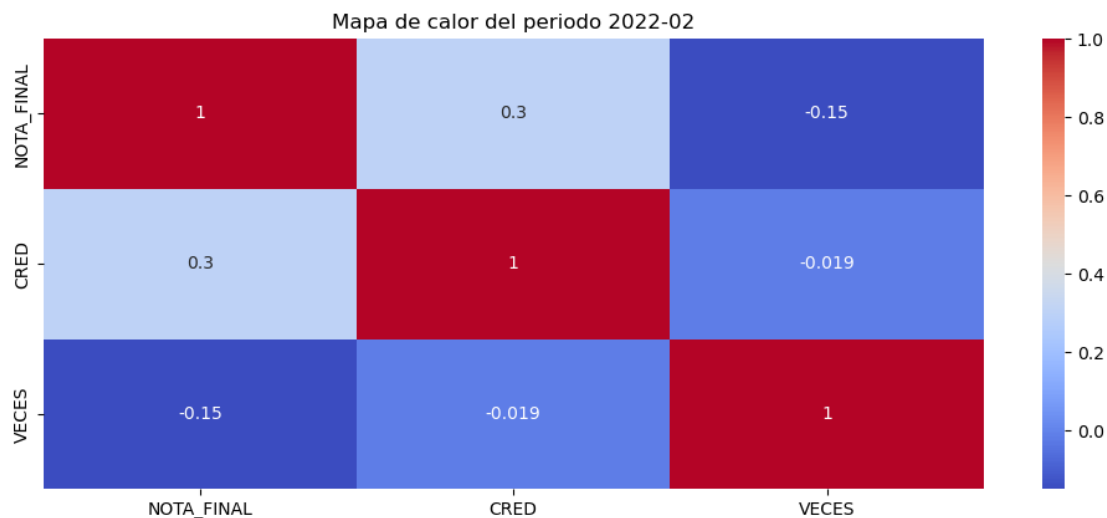


Ilustración 309

Mapa de calor del cuarto periodo de las variables NOTA_FINAL, CRED y VECES del reporte de Asignaturas Perdidas



A partir de los gráficos de calor de las tres variables, se puede determinar una correlación negativa débil en todos los periodos entre las calificaciones finales y la cantidad de veces que un estudiante repite una determinada asignatura, esto sugiere que, en general, los estudiantes que repiten una asignatura tienden a obtener calificaciones más bajas en comparación con aquellos que no la repiten; sin embargo, la correlación es bastante débil, lo que implica que existen otros factores que influyen en el rendimiento académico además de la repetición de asignaturas; así, pues, esta variable no resulta útil para la investigación.

Adicionalmente, se puede visualizar cómo la correlación entre las calificaciones finales y la cantidad de créditos de una determinada asignatura presenta una correlación positiva débil, sin embargo, al estar cerca de 0.5, es tangible la consideración de una relación.

Finalmente, se puede observar otra correlación negativa débil entre la cantidad de créditos de una asignatura y la cantidad de veces que se ha repetido la misma, sin embargo, en esta ocasión es más evidente la inexistencia de cualquier posible relación significativa.

En síntesis, y manera concluyente de la exploración de relaciones con respecto al reporte de asignaturas perdidas, con lo que respecta a las hipótesis:

- Las variables referidas al plan de estudios del estudiante, su facultad, carrera, nivel, créditos de la asignatura y área de la asignatura influyen en el rendimiento de la nota final de los estudiantes.
- Las variables referidas al plan de estudios del estudiante, su facultad, carrera, nivel, créditos de la asignatura y área de la asignatura influyen en la cantidad de veces que el estudiante ha repetido alguna asignatura.

Se pudo concretar que, se comprobó la primera hipótesis con excepción del uso de variables como PLAN_E y VECES, por otra parte, la segunda hipótesis resultó errónea, ya que, se demostró que la variable VECES no representa valor para el futuro modelo de machine learning.

En otras palabras, en conjunto con los análisis exploratorios de variables tanto numéricas como categóricas, y la presente exploración de relaciones, se pudo determinar la siguiente relevancia e irrelevancia de variables con respecto al reporte de asignaturas perdidas:

Tabla 18
Determinación de irrelevancia y relevancia de variables del reporte de Asignaturas Perdidas

Variable	Relevancia	Comentarios
IDD	No aplica	Esta variable es ideada para permitir una futura integración de datos.
PERIODO	No aplica	Esta variable es ideada para permitir una futura integración de datos.

PLAN_E	No relevante	No aplica
DESC_FACULTAD	Relevante	No aplica
DESC_PROGRAMA	Relevante	Es posible que deba ser condicionada.
NIVEL_ESTUDIANTE	Relevante	No aplica
MATE	Relevante	No aplica
NOTA_FINAL	No aplica	Esta variable tiene el objetivo de permitir el análisis de las variables que más afectan al rendimiento académico, pero, dado el enfoque de la investigación no resulta útil.
APROBADO	Relevante	Esta variable será la variable objetivo del modelo de machine learning.
VECES	No relevante	No aplica
CRED	Relevante	No aplica

En consecuencia, de haber finalizado la exploración de relaciones de las variables del reporte de asignaturas perdidas, es el momento apropiado de continuar la exploración con el reporte de calificaciones por carrera, para esto, Se tiene en cuenta las siguientes hipótesis a responder:

- Las variables referidas a la facultad, docentes, área de la asignatura y carrera influyen en el rendimiento de las notas finales de los estudiantes.

Por lo que, se comenzará indagando la relación entre las facultades a las que pertenecen los estudiantes y sus calificaciones finales a través de los diagramas de violín consiguientes:

Ilustración 310

Código que genera los gráficos para la relación entre facultad y nota_final del reporte de Calificaciones por Carrera

```
# Ordenamos las categorías alfabéticamente
orden_categorias = set()
for calificacion in calificaciones:
    orden_categorias.update(list(sorted(calificacion["facultad"].unique())))

# Grid para los gráficos
fig, axes = plt.subplots(4, 1, figsize=(10, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='facultad', y='nota_final', data=calificaciones[0], order=orden_categorias, ax=axes[0])
axes[0].set_xticklabels(axes[0].get_xticklabels(), rotation=90)

axes[1].set_title('Periodo 2021-02')
sns.violinplot(x='facultad', y='nota_final', data=calificaciones[1], order=orden_categorias, ax=axes[1])
axes[1].set_xticklabels(axes[1].get_xticklabels(), rotation=90)

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='facultad', y='nota_final', data=calificaciones[2], order=orden_categorias, ax=axes[2])
axes[2].set_xticklabels(axes[2].get_xticklabels(), rotation=90)

axes[3].set_title('Periodo 2022-02')
sns.violinplot(x='facultad', y='nota_final', data=calificaciones[3], order=orden_categorias, ax=axes[3])
axes[3].set_xticklabels(axes[3].get_xticklabels(), rotation=90)

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=2)

plt.show() # Mostrar los gráficos
```

Ilustración 311

Diagrama de Violín del primer periodo de la relación entre facultad y nota_final del reporte de Calificaciones por Carrera

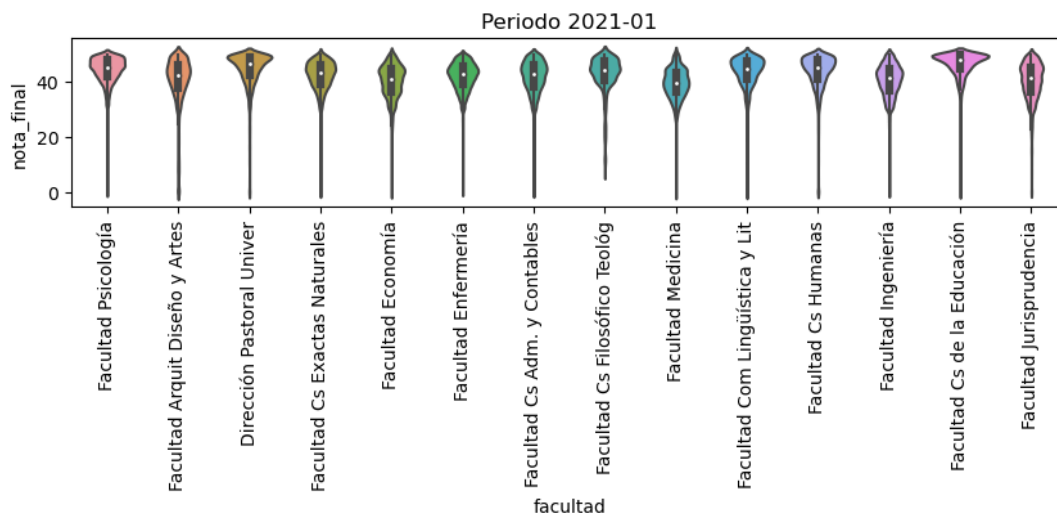


Ilustración 312

Diagrama de Violín del segundo periodo de la relación entre facultad y nota_final del reporte de Calificaciones por Carrera

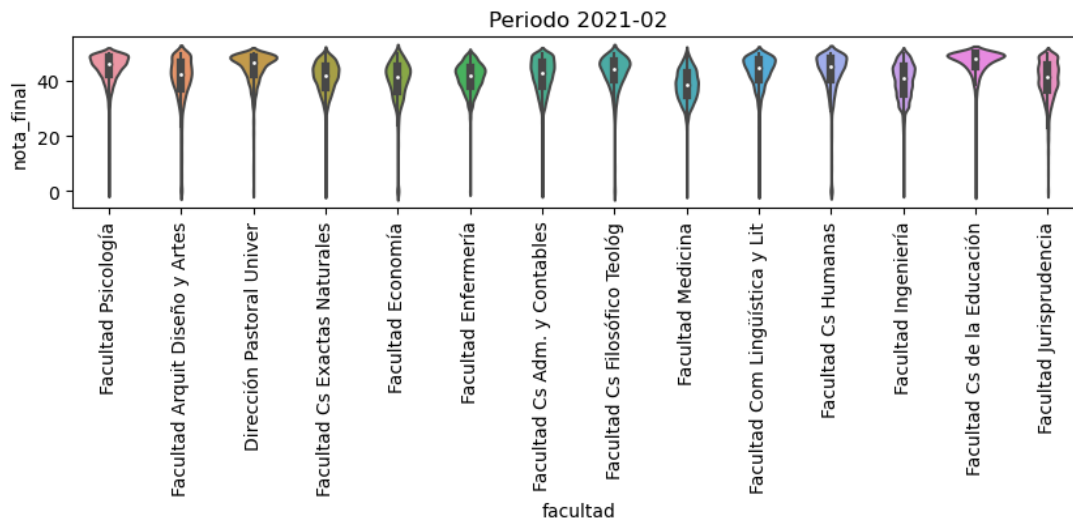


Ilustración 313

Diagrama de Violín del tercer periodo de la relación entre facultad y nota_final del reporte de Calificaciones por Carrera

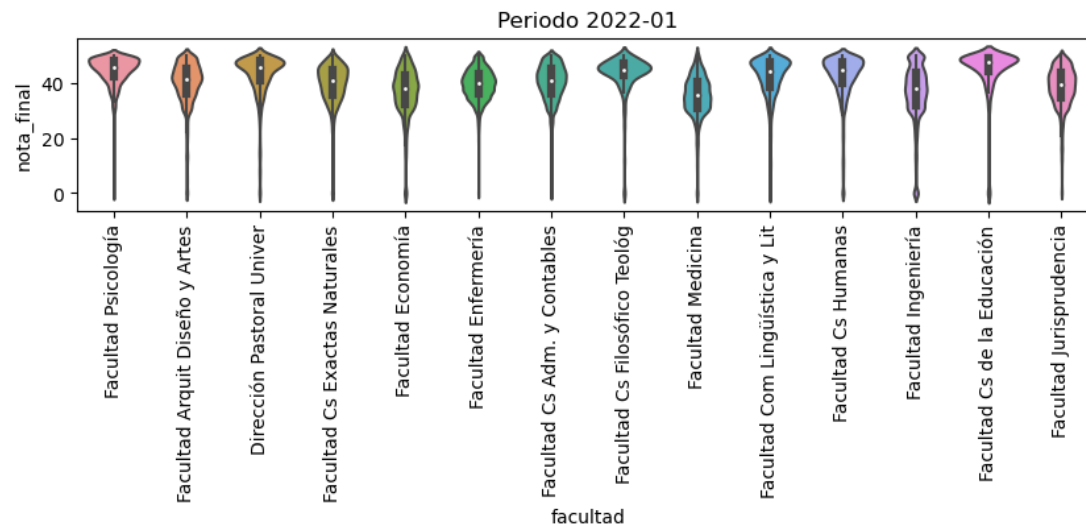
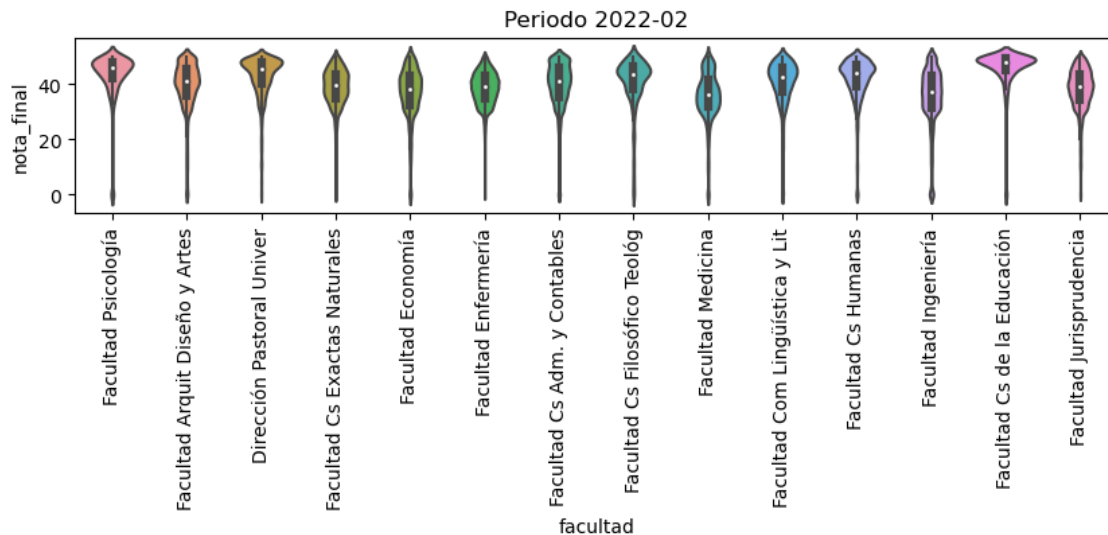


Ilustración 314

Diagrama de Violín del cuarto periodo de la relación entre facultad y nota_final del reporte de Calificaciones por Carrera



En conclusión, es evidente una estrecha relación entre las distribuciones de notas finales y las facultades, lo que resalta en que las facultades poseen procesos inherentes que influyen directamente en el rendimiento académico universitario.

Por otro lado, considerando la evaluación de la relación entre los docentes y las notas finales de los estudiantes, se debe tener en cuenta que, debido a la gran cantidad de datos que poseen los docentes no es factible explorar relaciones de la misma manera en la que se ha realizado hasta el momento; es decir, para llevar a cabo la exploración de esta relación se realizarán muestreos aleatorios, y a partir de estas muestras, se realizará la exploración de relaciones.

Por consiguiente, los resultados del primer muestreo son:

Ilustración 315

Código que genera los gráficos del primer muestreo para la relación entre id_banner_docente y nota_final del reporte de Calificaciones por Carrera

```
# Se crean dataframes temporales para evitar problemas en los dataframes originales
dataframes_temporales = []
for calificacion in calificaciones:
    dataframes_temporales.append(calificacion.copy())

# Se establece el tamaño de la muestra y el factor de pseudo aleatoriedad
sample_n = 4
random_state = 1

# Tomamos muestras aleatorias
for i in range(len(dataframes_temporales)):
    dataframes_temporales[i] = dataframes_temporales[i].sample(n=sample_n, random_state=random_state)

# Ordenamos las categorías de docentes alfabéticamente
orden_categorias = set()
for temporal in dataframes_temporales:
    orden_categorias.update(list(sorted(temporal["id_banner_docente"].unique())))

# Grid para los gráficos
fig, axes = plt.subplots(4, 1, figsize=(15, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='id_banner_docente', y='nota_final', data=calificaciones[0], order=orden_categorias, ax=axes[0])
axes[0].set_xticklabels(axes[0].get_xticklabels(), rotation=90)

axes[1].set_title('Periodo 2021-02')
sns.violinplot(x='id_banner_docente', y='nota_final', data=calificaciones[1], order=orden_categorias, ax=axes[1])
axes[1].set_xticklabels(axes[1].get_xticklabels(), rotation=90)

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='id_banner_docente', y='nota_final', data=calificaciones[2], order=orden_categorias, ax=axes[2])
axes[2].set_xticklabels(axes[2].get_xticklabels(), rotation=90)

axes[3].set_title('Periodo 2022-02')
sns.violinplot(x='id_banner_docente', y='nota_final', data=calificaciones[3], order=orden_categorias, ax=axes[3])
axes[3].set_xticklabels(axes[3].get_xticklabels(), rotation=90)

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.4)

plt.show() # Mostrar los gráficos
```

Ilustración 316

Diagrama de Violín del primer muestreo del primer periodo de la relación entre id_banner_docente y nota_final del reporte de Calificaciones por Carrera

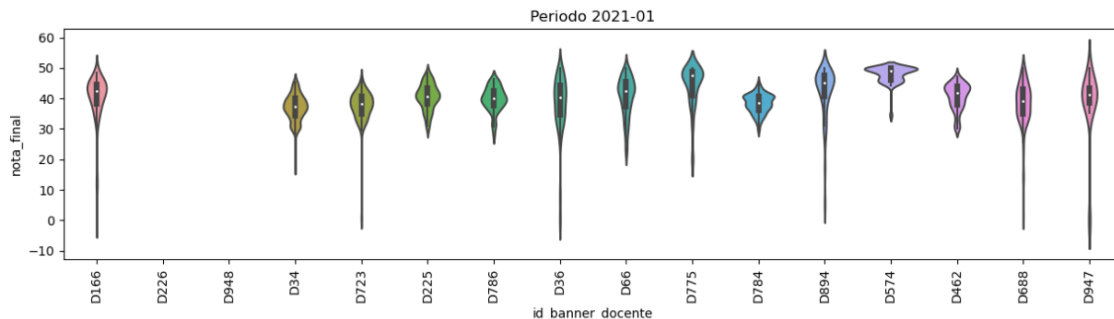


Ilustración 317

Diagrama de Violín del primer muestreo del segundo periodo de la relación entre `id_banner_docente` y `nota_final` del reporte de Calificaciones por Carrera

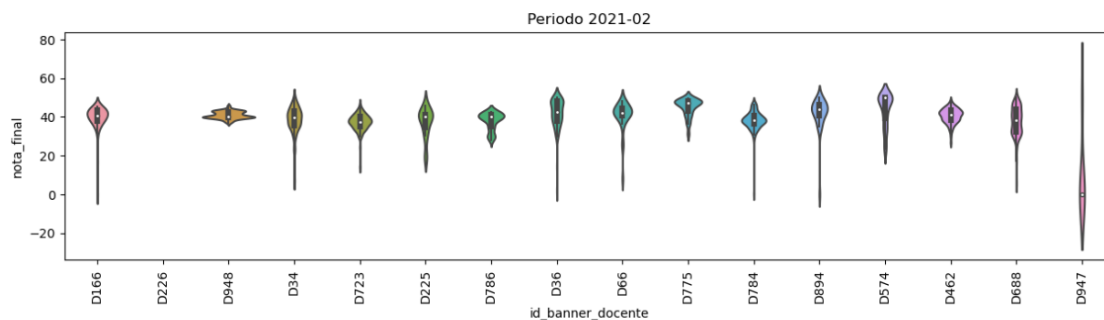


Ilustración 318

Diagrama de Violín del primer muestreo del tercer periodo de la relación entre `id_banner_docente` y `nota_final` del reporte de Calificaciones por Carrera

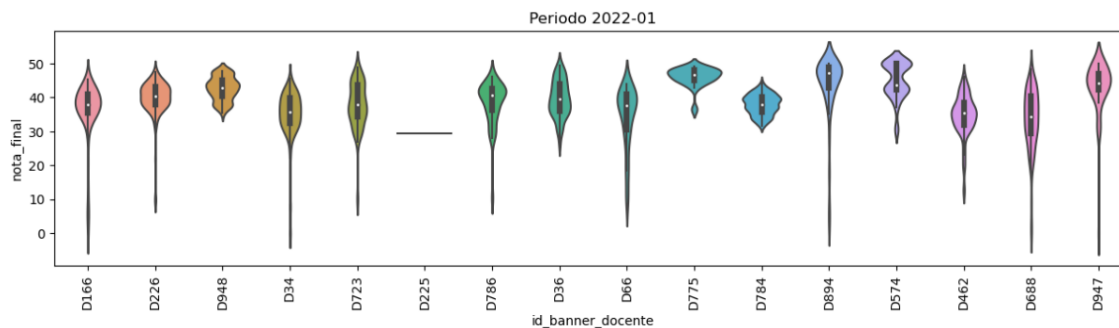
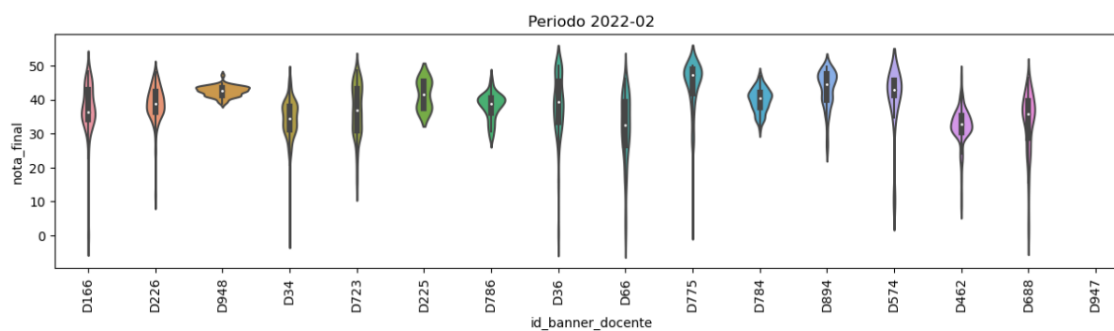


Ilustración 319

Diagrama de Violín del primer muestreo del cuarto periodo de la relación entre `id_banner_docente` y `nota_final` del reporte de Calificaciones por Carrera



En segundo término, los resultados del segundo muestreo aleatorio son:

Ilustración 320

Código que genera los gráficos del segundo muestreo para la relación entre id_banner_docente y nota_final del reporte de Calificaciones por Carrera

```
# Se crean dataframes temporales para evitar problemas en los dataframes originales
dataframes_temporales = []
for calificacion in calificaciones:
    dataframes_temporales.append(calificacion.copy())

# Se establece el tamaño de la muestra y el factor de pseudo aleatoriedad
sample_n = 4
random_state = 10

# Tomamos muestras aleatorias
for i in range(len(dataframes_temporales)):
    dataframes_temporales[i] = dataframes_temporales[i].sample(n=sample_n, random_state=random_state)

# Ordenamos las categorías de docentes alfabéticamente
orden_categorias = set()
for temporal in dataframes_temporales:
    orden_categorias.update(list(sorted(temporal["id_banner_docente"].unique())))

# Grid para los gráficos
fig, axes = plt.subplots(4, 1, figsize=(15, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='id_banner_docente', y='nota_final', data=calificaciones[0], order=orden_categorias, ax=axes[0])
axes[0].set_xticklabels(axes[0].get_xticklabels(), rotation=90)

axes[1].set_title('Periodo 2021-02')
sns.violinplot(x='id_banner_docente', y='nota_final', data=calificaciones[1], order=orden_categorias, ax=axes[1])
axes[1].set_xticklabels(axes[1].get_xticklabels(), rotation=90)

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='id_banner_docente', y='nota_final', data=calificaciones[2], order=orden_categorias, ax=axes[2])
axes[2].set_xticklabels(axes[2].get_xticklabels(), rotation=90)

axes[3].set_title('Periodo 2022-02')
sns.violinplot(x='id_banner_docente', y='nota_final', data=calificaciones[3], order=orden_categorias, ax=axes[3])
axes[3].set_xticklabels(axes[3].get_xticklabels(), rotation=90)

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.4)

plt.show() # Mostrar los gráficos
```

Ilustración 321

Diagrama de Violín del segundo muestreo del primer periodo de la relación entre id_banner_docente y nota_final del reporte de Calificaciones por Carrera

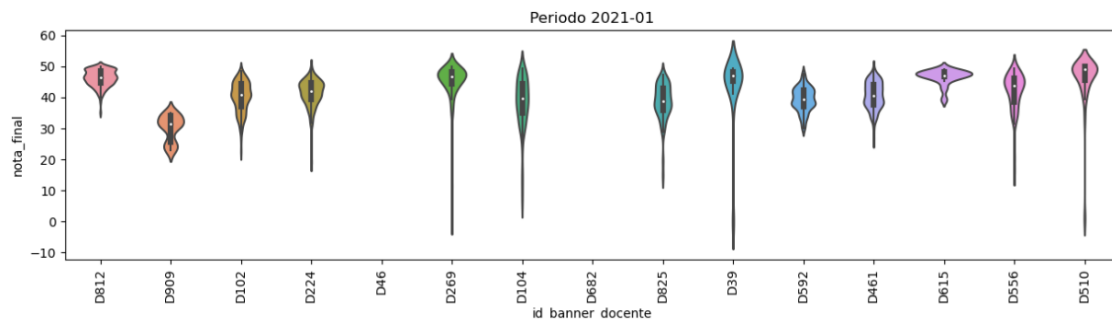


Ilustración 322

Diagrama de Violín del segundo muestreo del segundo periodo de la relación entre `id_banner_docente` y `nota_final` del reporte de Calificaciones por Carrera

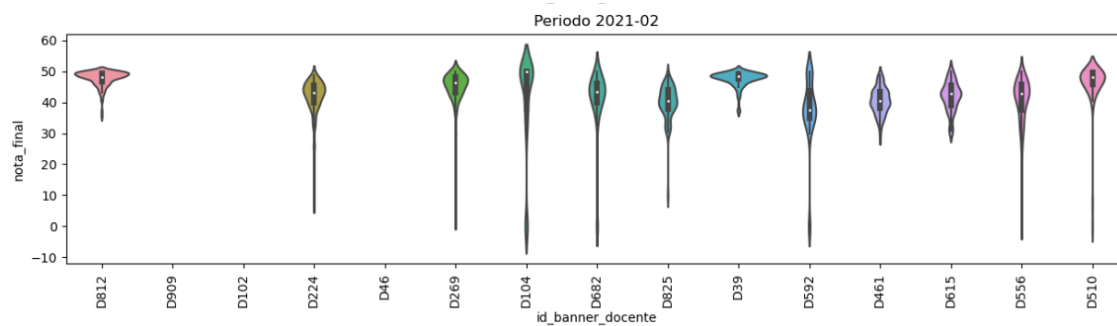


Ilustración 323

Diagrama de Violín del segundo muestreo del tercer periodo de la relación entre `id_banner_docente` y `nota_final` del reporte de Calificaciones por Carrera

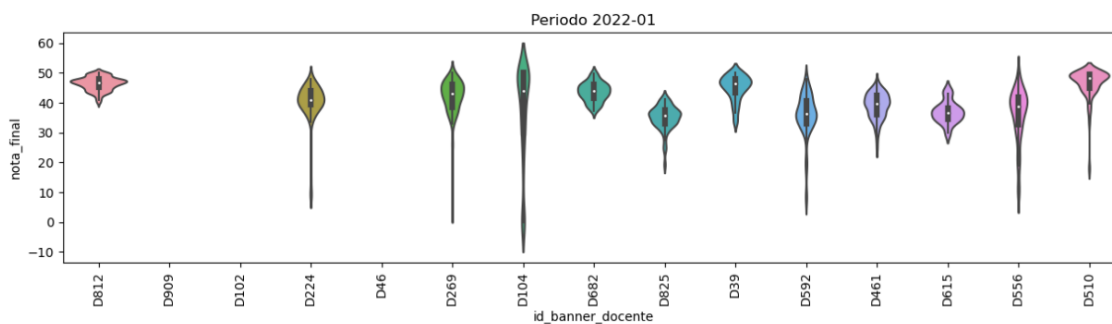
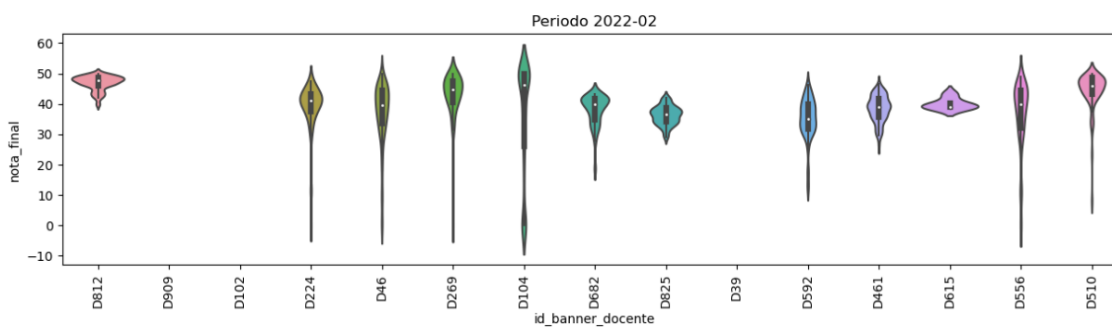


Ilustración 324

Diagrama de Violín del segundo muestreo del cuarto periodo de la relación entre `id_banner_docente` y `nota_final` del reporte de Calificaciones por Carrera



Finalmente, los resultados del tercer muestreo aleatorio de los docentes son:

Ilustración 325

Código que genera los gráficos del tercer muestreo para la relación entre id_banner_docente y nota_final del reporte de Calificaciones por Carrera

```
# Se crean dataframes temporales para evitar problemas en los dataframes originales
dataframes_temporales = []
for calificacion in calificaciones:
    dataframes_temporales.append(calificacion.copy())

# Se establece el tamaño de la muestra y el factor de pseudo aleatoriedad
sample_n = 4
random_state = 20

# Tomamos muestras aleatorias
for i in range(len(dataframes_temporales)):
    dataframes_temporales[i] = dataframes_temporales[i].sample(n=sample_n, random_state=random_state)

# Ordenamos las categorías de docentes alfabéticamente
orden_categorias = set()
for temporal in dataframes_temporales:
    orden_categorias.update(list(sorted(temporal["id_banner_docente"].unique())))

# Grid para los gráficos
fig, axes = plt.subplots(4, 1, figsize=(15, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='id_banner_docente', y='nota_final', data=calificaciones[0], order=orden_categorias, ax=axes[0])
axes[0].set_xticklabels(axes[0].get_xticklabels(), rotation=90)

axes[1].set_title('Periodo 2021-02')
sns.violinplot(x='id_banner_docente', y='nota_final', data=calificaciones[1], order=orden_categorias, ax=axes[1])
axes[1].set_xticklabels(axes[1].get_xticklabels(), rotation=90)

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='id_banner_docente', y='nota_final', data=calificaciones[2], order=orden_categorias, ax=axes[2])
axes[2].set_xticklabels(axes[2].get_xticklabels(), rotation=90)

axes[3].set_title('Periodo 2022-02')
sns.violinplot(x='id_banner_docente', y='nota_final', data=calificaciones[3], order=orden_categorias, ax=axes[3])
axes[3].set_xticklabels(axes[3].get_xticklabels(), rotation=90)

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.4)

plt.show() # Mostrar los gráficos
```

Ilustración 326

Diagrama de Violín del tercer muestreo del primer periodo de la relación entre id_banner_docente y nota_final del reporte de Calificaciones por Carrera

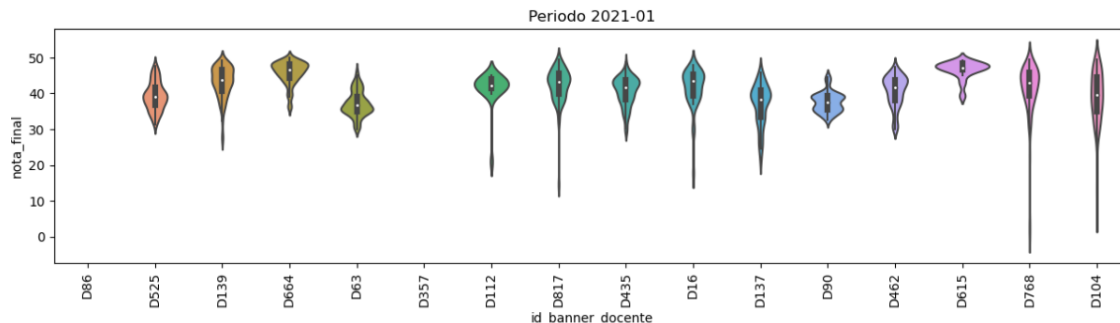


Ilustración 327

Diagrama de Violín del tercer muestreo del segundo periodo de la relación entre id_banner_docente y nota_final del reporte de Calificaciones por Carrera

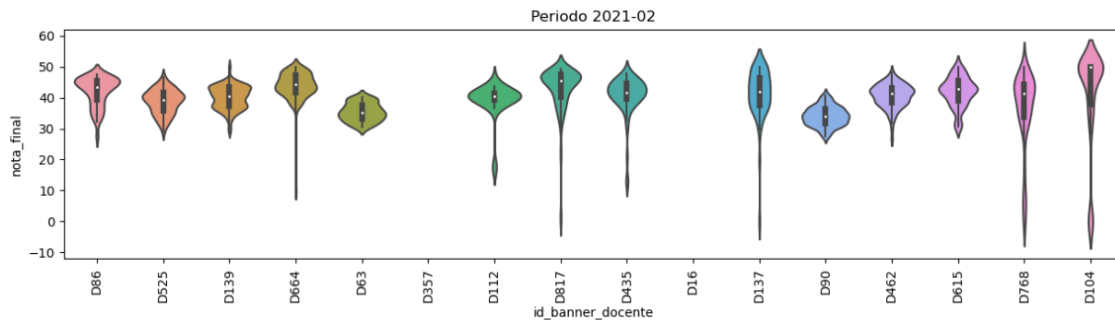


Ilustración 328

Diagrama de Violín del tercer muestreo del tercer periodo de la relación entre id_banner_docente y nota_final del reporte de Calificaciones por Carrera

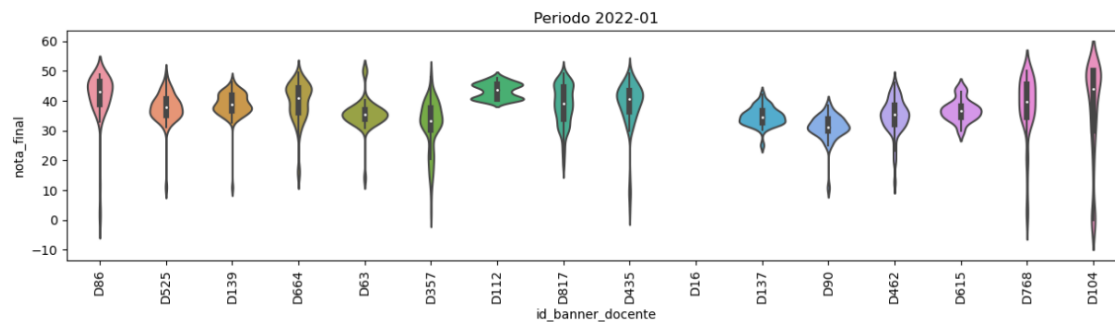
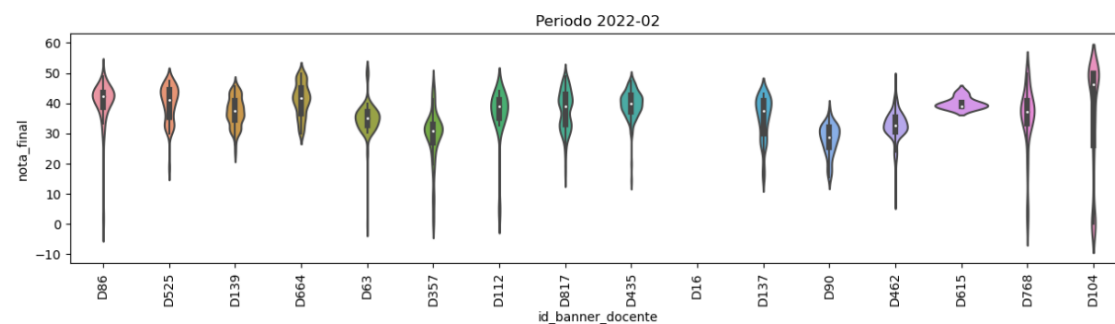


Ilustración 329

Diagrama de Violín del tercer muestreo del cuarto periodo de la relación entre id_banner_docente y nota_final del reporte de Calificaciones por Carrera



En conclusión de la relación entre los docentes y el rendimiento académico de los estudiantes, a partir de los análisis con muestreos aleatorios diferentes, se puede concretar que existe una relación entre el docente y el desempeño académico de sus estudiantes, pues, dentro de las muestras analizadas se puede observar que el patrón de la figura de las distribuciones de

calificaciones finales de estudiantes tiene poca variabilidad a través de los periodos con el mismo docente, y estos son diferentes entre docentes. Esto se respalda por una extensa investigación que menciona la influencia de los docentes en el rendimiento estudiantil. (Guizado Salazar, Valenzuela Rodríguez, & Vallejo Quispe, 2020, pág. 3) Así como también se habla de cómo la relación entre docente y estudiante influye en su rendimiento. (Educación 3.0, s.f.). Por lo tanto, se puede concluir de manera determinante que el uso de esta variable es crucial para predecir el rendimiento académico.

En diferente sentido, la relación entre las áreas de asignaturas y las notas finales de los estudiantes se evaluará mediante diagramas de violín, sus resultados podrán ser apreciados a continuación:

Ilustración 330

Código que genera los gráficos para la relación entre `descripcion_materia` y `nota_final` del reporte de Calificaciones por Carrera

```
# Ordenamos las categorías alfabéticamente
orden_categorias = set()
for calificacion in calificaciones:
    orden_categorias.update(list(sorted(calificacion["descripcion_materia"].unique())))

# Grid para los gráficos
fig, axes = plt.subplots(4, 1, figsize=(15, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='descripcion_materia', y='nota_final', data=calificaciones[0], order=orden_categorias, ax=axes[0])
axes[0].set_xticklabels(axes[0].get_xticklabels(), rotation=90)

axes[1].set_title('Periodo 2021-02')
sns.violinplot(x='descripcion_materia', y='nota_final', data=calificaciones[1], order=orden_categorias, ax=axes[1])
axes[1].set_xticklabels(axes[1].get_xticklabels(), rotation=90)

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='descripcion_materia', y='nota_final', data=calificaciones[2], order=orden_categorias, ax=axes[2])
axes[2].set_xticklabels(axes[2].get_xticklabels(), rotation=90)

axes[3].set_title('Periodo 2022-02')
sns.violinplot(x='descripcion_materia', y='nota_final', data=calificaciones[3], order=orden_categorias, ax=axes[3])
axes[3].set_xticklabels(axes[3].get_xticklabels(), rotation=90)

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=2.4)

plt.show() # Mostrar los gráficos
```

Ilustración 331

Diagrama de Violín del primer periodo de la relación entre descripcion_materia y nota_final del reporte de Calificaciones por Carrera

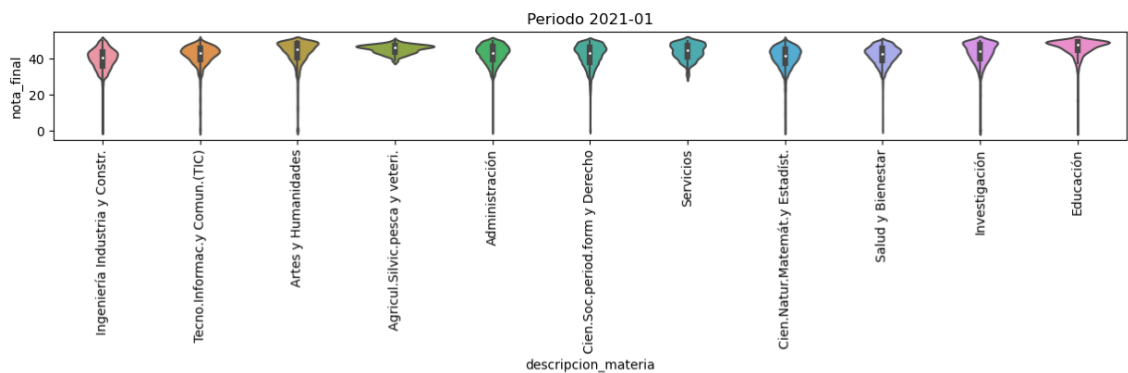


Ilustración 332

Diagrama de Violín del segundo periodo de la relación entre descripcion_materia y nota_final del reporte de Calificaciones por Carrera

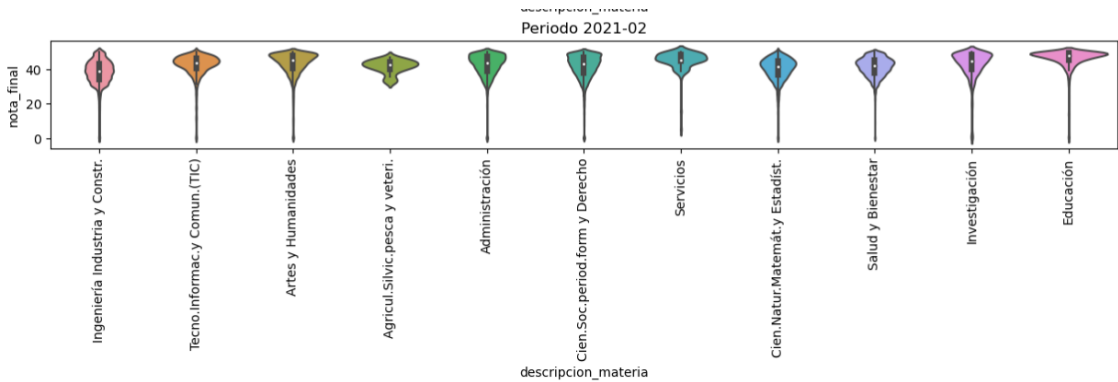


Ilustración 333

Diagrama de Violín del tercer periodo de la relación entre descripcion_materia y nota_final del reporte de Calificaciones por Carrera

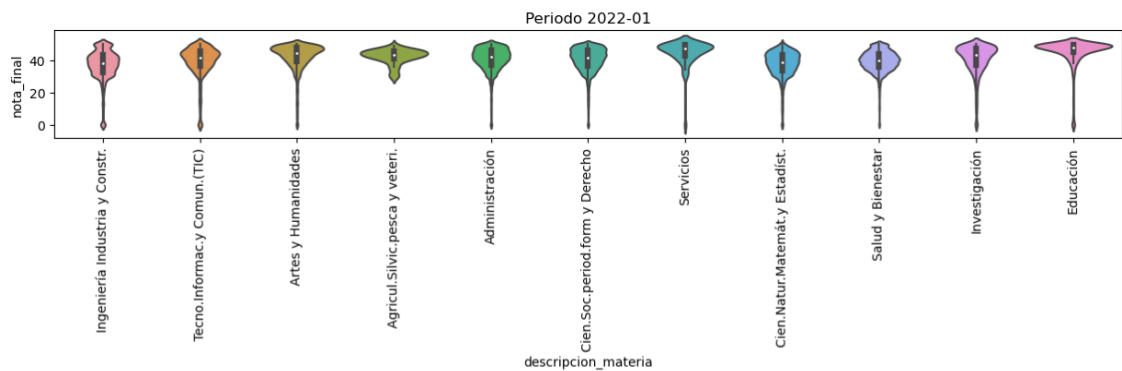
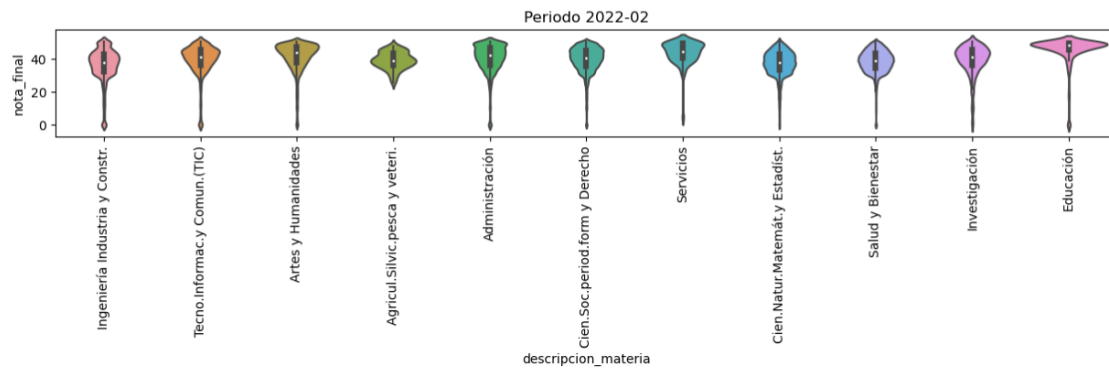


Ilustración 334

Diagrama de Violín del cuarto periodo de la relación entre descripcion_materia y nota_final del reporte de Calificaciones por Carrera



En conclusión, así como se comprobó la relación entre las calificaciones finales y las áreas de asignaturas en el reporte de asignaturas perdidas, en este caso igualmente se evidencia una estrecha relación entre el área de la asignatura cursada y el desempeño en la misma.

Similar a un análisis realizado para el reporte de asignaturas perdidas, se realizará un análisis de la posible relación entre las carreras y las notas finales de los estudiantes en el reporte de calificaciones por carrera, sus resultados son detallados a continuación:

Ilustración 335

Código que genera los gráficos para la relación entre descripcion_programa y nota_final del reporte de Calificaciones por Carrera

```
# Ordenamos Las categorías alfabéticamente
orden_categorias = set()
for calificacion in calificaciones:
    orden_categorias.update(list(sorted(calificacion["descripcion_programa"].unique())))

# Grid para Los gráficos
fig, axes = plt.subplots(4, 1, figsize=(15, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.boxplot(x='descripcion_programa', y='nota_final', data=calificaciones[0], order=orden_categorias, ax=axes[0])
axes[0].set_xticklabels(axes[0].get_xticklabels(), rotation=90)

axes[1].set_title('Periodo 2021-02')
sns.boxplot(x='descripcion_programa', y='nota_final', data=calificaciones[1], order=orden_categorias, ax=axes[1])
axes[1].set_xticklabels(axes[1].get_xticklabels(), rotation=90)

axes[2].set_title('Periodo 2022-01')
sns.boxplot(x='descripcion_programa', y='nota_final', data=calificaciones[2], order=orden_categorias, ax=axes[2])
axes[2].set_xticklabels(axes[2].get_xticklabels(), rotation=90)

axes[3].set_title('Periodo 2022-02')
sns.boxplot(x='descripcion_programa', y='nota_final', data=calificaciones[3], order=orden_categorias, ax=axes[3])
axes[3].set_xticklabels(axes[3].get_xticklabels(), rotation=90)

# Ajustar el espacio entre Los subplots
plt.subplots_adjust(wspace=0.2, hspace=2.4)

plt.show() # Mostrar Los gráficos
```

Ilustración 336

Diagrama de Caja y Bigotes del primer periodo de la relación entre descripcion_programa y nota_final del reporte de Calificaciones por Carrera

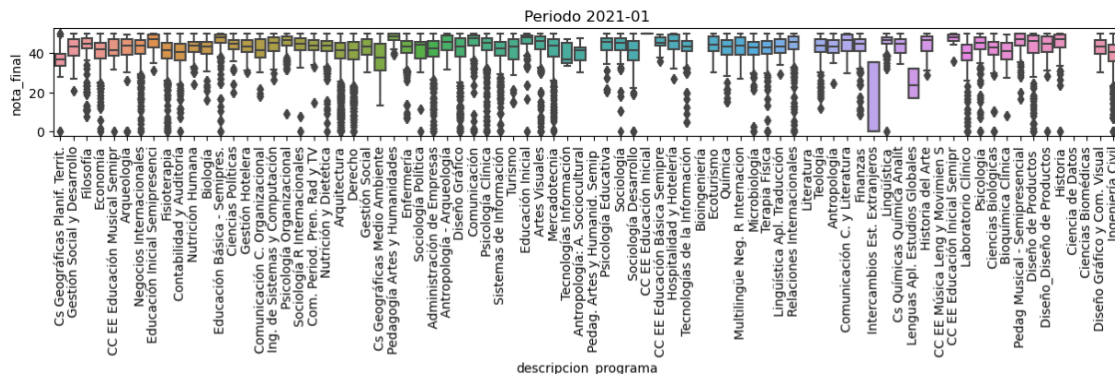


Ilustración 337

Diagrama de Caja y Bigotes del segundo periodo de la relación entre descripcion_programa y nota_final del reporte de Calificaciones por Carrera

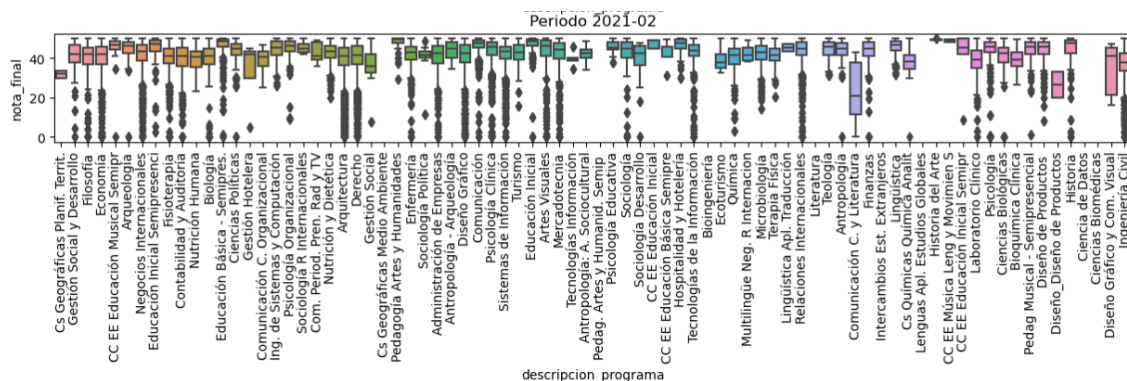


Ilustración 338

Diagrama de Caja y Bigotes del tercer periodo de la relación entre descripcion_programa y nota_final del reporte de Calificaciones por Carrera

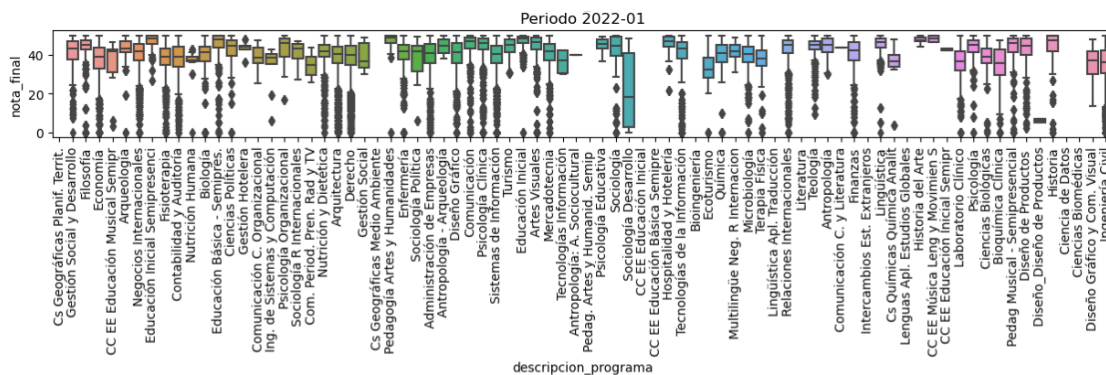
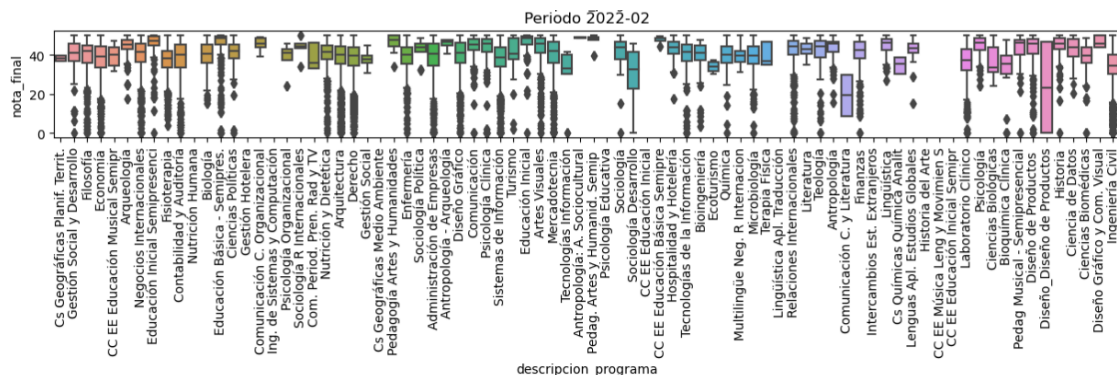


Ilustración 339

Diagrama de Caja y Bigotes del cuarto periodo de la relación entre descripcion_programa y nota_final del reporte de Calificaciones por Carrera



En conclusión, se puede observar que, en general, existe una relación directa entre el rendimiento académico y las calificaciones finales en todas las carreras, esto se evidencia en el

hecho de que la distribución de calificaciones no varía considerablemente a lo largo de los diferentes periodos dentro de una misma carrera, mientras que, sí existen variaciones significativas entre las diferentes carreras; sin embargo, es notorio que existen varias carreras con un aproximado a 0 registros, por lo que se debe considerar en la fase de preparación de los datos.

Finalmente se concluido la exploración de las relaciones en el reporte de calificaciones por carrera, sin embargo, para conocer sus resultados se debe recalcar que la hipótesis que se cuestionó fue realizada en base a variables referentes de rendimiento académico, en esta ocasión exclusivamente se realizó en base a las notas finales de los estudiantes por las razones descritas al comienzo de este subcapítulo, por lo que, la hipótesis fue:

- Las variables referidas a la facultad, docente, área de la asignatura y carrera influyen en el rendimiento de las notas finales de un estudiante.

Y, tras los análisis realizados, se pudo comprobar satisfactoriamente la hipótesis, aunque, con algunas excepciones; de todas maneras, a partir de sus resultados y de todos los resultados de todos los análisis llevados a cabo sobre el reporte de calificaciones por carrera, se puede concluir que se han identificado a detalle variables relevantes e irrelevante, las mismas serán descritas en la siguiente tabla:

Tabla 19
Determinación de irrelevancia y relevancia de variables del reporte de Calificaciones por Carrera

Variable	Relevancia	Comentarios
id_banner_estudiante	No aplica	Esta variable es ideada para permitir una futura integración de datos.
codigo_periodes	No aplica	Esta variable es ideada para permitir una futura integración de datos.
facultad	Relevante	

nota_parcial	Irrelevante	Esta explicación se puede encontrar en el capítulo 8 de este cuaderno.
nota_final	No aplica	Esta variable tiene el objetivo de permitir el análisis de las variables que más afectan al rendimiento académico, pero, dado el enfoque de la investigación no resulta útil.
id_banner_docente	Relevante	
descripcion_materia	Relevante	
descripcion_programa	Relevante	
nota_final_SHATCKG	No relevante	Esta variable fue concluida como no relevante en el análisis de distribuciones del mismo reporte, es por ello por lo que no se realizaron análisis de relaciones con esta variable.

Una vez concretada la identificación de las distintas variables relevantes e irrelevantes del reporte de calificaciones por carrera, es momento de encaminarse al último paso de la exploración de relaciones, la exploración en el reporte de becas, sin embargo, antes de comenzar es prudente recordar al distinguido lector que las hipótesis que se tratarán de resolver son:

- Las variables referidas al género, etnia, tipo de colegio, carrera, nivel del estudiante, nivel de asignaturas tomadas, categoría de beca socioeconómica influyen en el rendimiento académico con respecto a la cantidad de materias en segunda matrícula.
- Las variables referidas al género, etnia, tipo de colegio, carrera, nivel del estudiante, nivel de asignaturas tomadas, categoría de beca socioeconómica y cantidad de asignaturas con segunda matrícula influyen en el rendimiento académico con respecto a la cantidad de materias en tercera matrícula.

A partir del recordatorio, se procederá a evaluar la relación entre el género de un estudiante y su cantidad de asignaturas con segunda matrícula:

Ilustración 340

Código que genera los gráficos para la relación entre GENERO y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas

```
# Grid para Los gráficos
fig, axes = plt.subplots(4, 1, figsize=(15, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='GENERO', y='NUM_MATERIAS_SEGUNDA_MATRICULA', data=becas[0], ax=axes[0])
axes[0].set_xticklabels(axes[0].get_xticklabels(), rotation=90)

axes[1].set_title('Periodo 2021-02')
sns.violinplot(x='GENERO', y='NUM_MATERIAS_SEGUNDA_MATRICULA', data=becas[1], ax=axes[1])
axes[1].set_xticklabels(axes[1].get_xticklabels(), rotation=90)

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='GENERO', y='NUM_MATERIAS_SEGUNDA_MATRICULA', data=becas[2], ax=axes[2])
axes[2].set_xticklabels(axes[2].get_xticklabels(), rotation=90)

axes[3].set_title('Periodo 2022-02')
sns.violinplot(x='GENERO', y='NUM_MATERIAS_SEGUNDA_MATRICULA', data=becas[3], ax=axes[3])
axes[3].set_xticklabels(axes[3].get_xticklabels(), rotation=90)

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.8)

plt.show() # Mostrar Los gráficos
```

Ilustración 341

Diagrama de Violín del primer periodo de la relación entre GENERO y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas

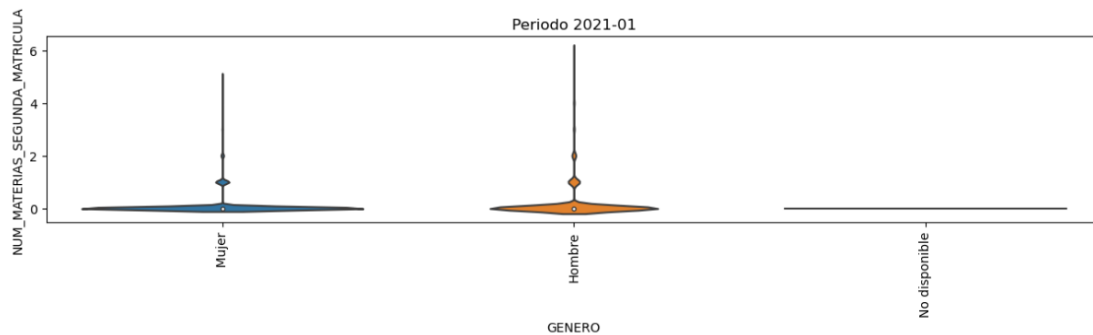


Ilustración 342

Diagrama de Violín del segundo periodo de la relación entre GENERO y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas

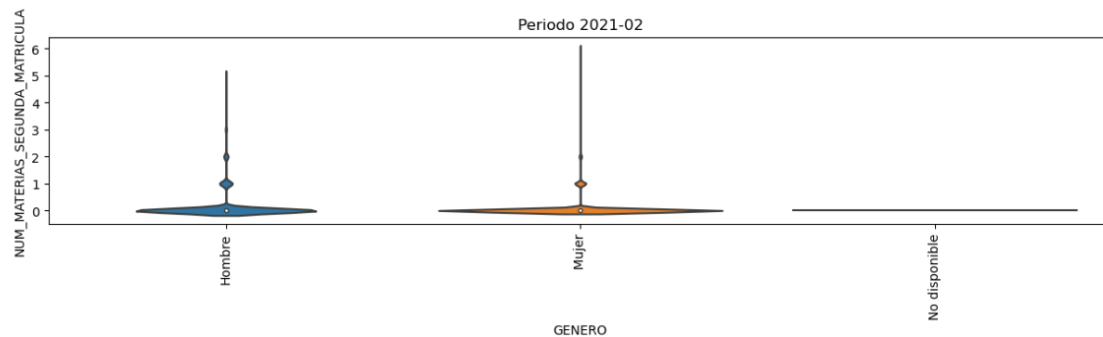


Ilustración 343

Diagrama de Violín del tercer periodo de la relación entre GENERO y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas

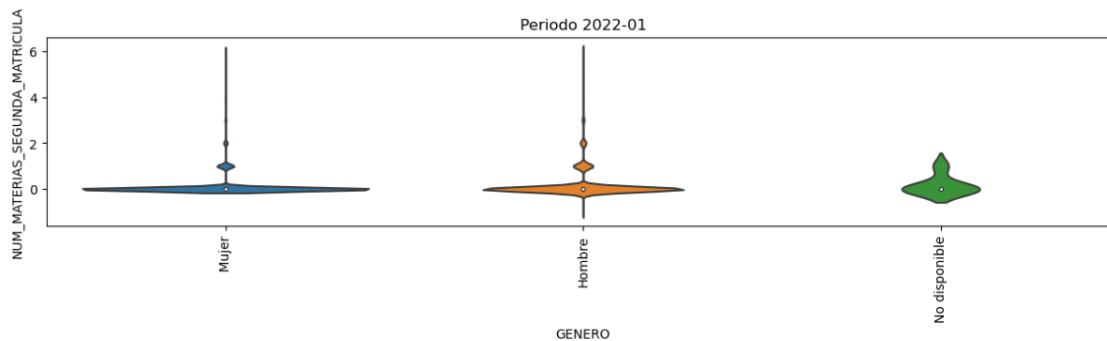
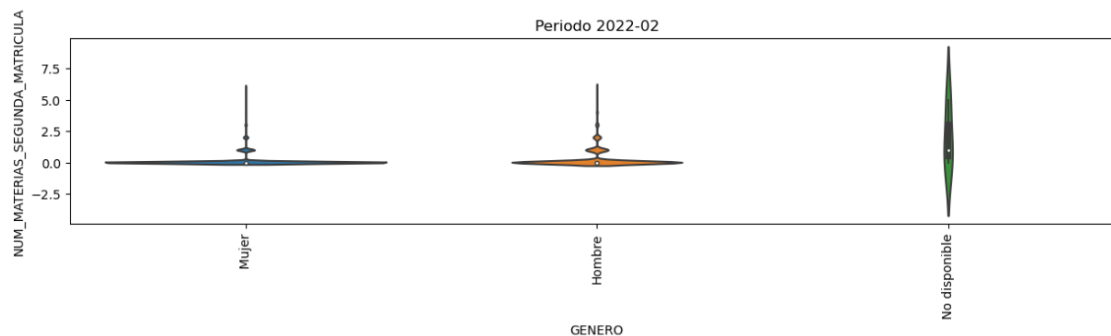


Ilustración 344

Diagrama de Violín del cuarto periodo de la relación entre GENERO y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas



En conclusión, se puede observar que no existe una relación clara entre el género de un estudiante y su cantidad de asignaturas con segunda matrícula. Por otra parte, considerando la

evaluación sobre la relación entre la etnia de un estudiante y su cantidad de asignaturas con segunda matrícula se obtuvieron los siguientes resultados:

Ilustración 345

Código que genera los gráficos para la relación entre ETNIA y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas

```
# Grid para Los gráficos
fig, axes = plt.subplots(4, 1, figsize=(15, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='ETNIA', y='NUM_MATERIAS_SEGUNDA_MATRICULA', data=becas[0], ax=axes[0])
axes[0].set_xticklabels(axes[0].get_xticklabels(), rotation=90)

axes[1].set_title('Periodo 2021-02')
sns.violinplot(x='ETNIA', y='NUM_MATERIAS_SEGUNDA_MATRICULA', data=becas[1], ax=axes[1])
axes[1].set_xticklabels(axes[1].get_xticklabels(), rotation=90)

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='ETNIA', y='NUM_MATERIAS_SEGUNDA_MATRICULA', data=becas[2], ax=axes[2])
axes[2].set_xticklabels(axes[2].get_xticklabels(), rotation=90)

axes[3].set_title('Periodo 2022-02')
sns.violinplot(x='ETNIA', y='NUM_MATERIAS_SEGUNDA_MATRICULA', data=becas[3], ax=axes[3])
axes[3].set_xticklabels(axes[3].get_xticklabels(), rotation=90)

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.8)

plt.show() # Mostrar Los gráficos
```

Ilustración 346

Diagrama de Violín del primer periodo de la relación entre ETNIA y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas

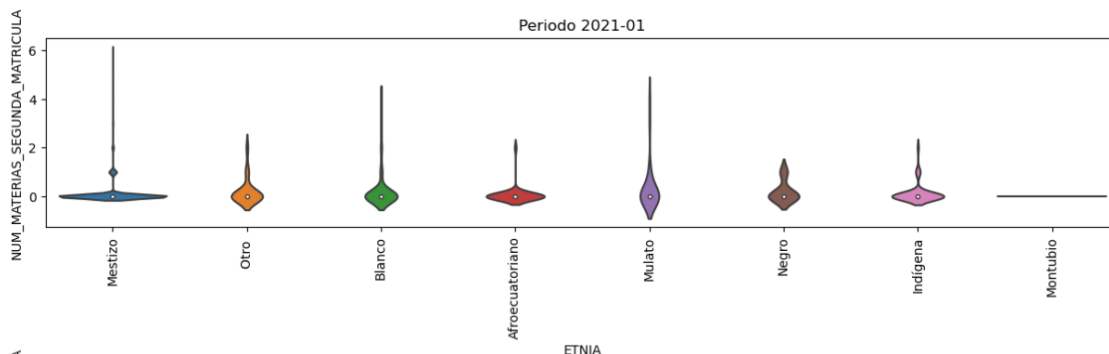


Ilustración 347

Diagrama de Violín del segundo periodo de la relación entre ETNIA y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas

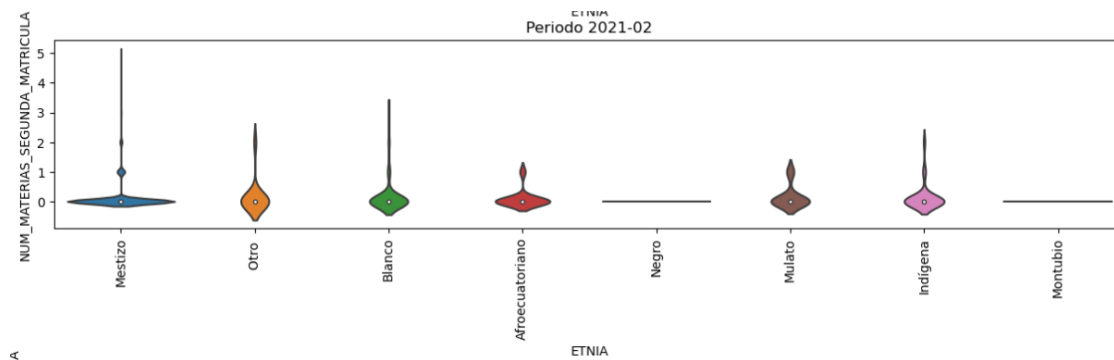


Ilustración 348

Diagrama de Violín del tercer periodo de la relación entre ETNIA y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas

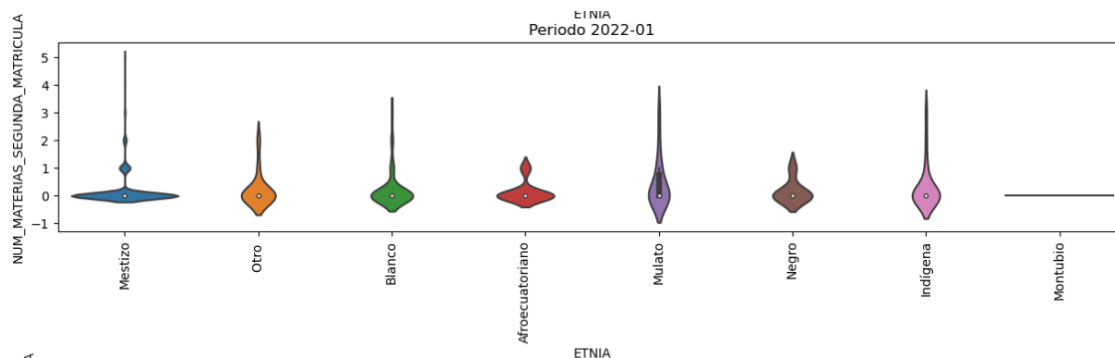
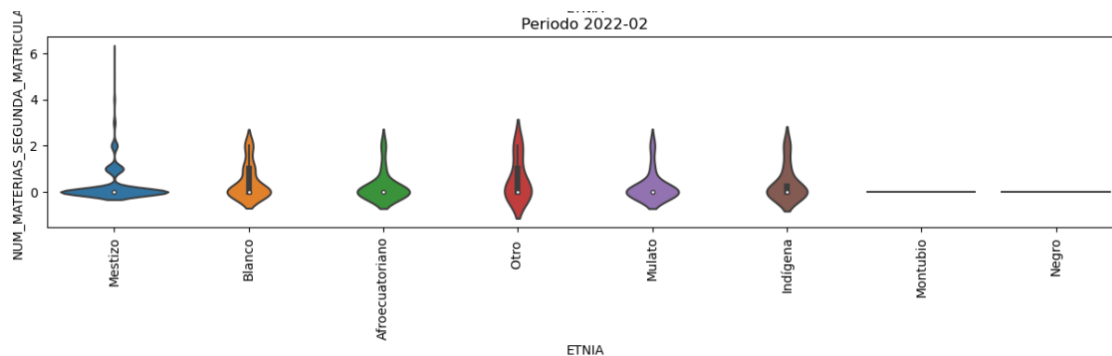


Ilustración 349

Diagrama de Violín del cuarto periodo de la relación entre ETNIA y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas



En conclusión, se puede apreciar que, aunque en ciertos períodos la relación entre la etnia de un estudiante y la cantidad de asignaturas con segunda matrícula que posee no sea tan evidente, se puede concluir que sí existe dicha relación; esto se puede deducir al observar el

patrón de figura que generan las distribuciones por cada etnia a través de los periodos académicos registrados y entre otras etnias. Sin embargo, también es notorio que el último período es el que se muestra más reacio con respecto a la existencia de una relación entre estas dos variables.

Como aspecto diferente, la evaluación de la relación entre el tipo de colegio al que perteneció un estudiante y su cantidad de asignaturas con segunda matrícula dio como resultado lo siguiente:

Ilustración 350

Código que genera los gráficos para la relación entre TIPO_COLEGIO y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas

```
# Grid para los gráficos
fig, axes = plt.subplots(4, 1, figsize=(15, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='TIPO_COLEGIO', y='NUM_MATERIAS_SEGUNDA_MATRICULA', data=becas[0], ax=axes[0])
axes[0].set_xticklabels(axes[0].get_xticklabels(), rotation=90)

axes[1].set_title('Periodo 2021-02')
sns.violinplot(x='TIPO_COLEGIO', y='NUM_MATERIAS_SEGUNDA_MATRICULA', data=becas[1], ax=axes[1])
axes[1].set_xticklabels(axes[1].get_xticklabels(), rotation=90)

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='TIPO_COLEGIO', y='NUM_MATERIAS_SEGUNDA_MATRICULA', data=becas[2], ax=axes[2])
axes[2].set_xticklabels(axes[2].get_xticklabels(), rotation=90)

axes[3].set_title('Periodo 2022-02')
sns.violinplot(x='TIPO_COLEGIO', y='NUM_MATERIAS_SEGUNDA_MATRICULA', data=becas[3], ax=axes[3])
axes[3].set_xticklabels(axes[3].get_xticklabels(), rotation=90)

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.8)

plt.show() # Mostrar los gráficos
```

Ilustración 351

Diagrama de Violín del primer periodo de la relación entre TIPO_COLEGIO y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas

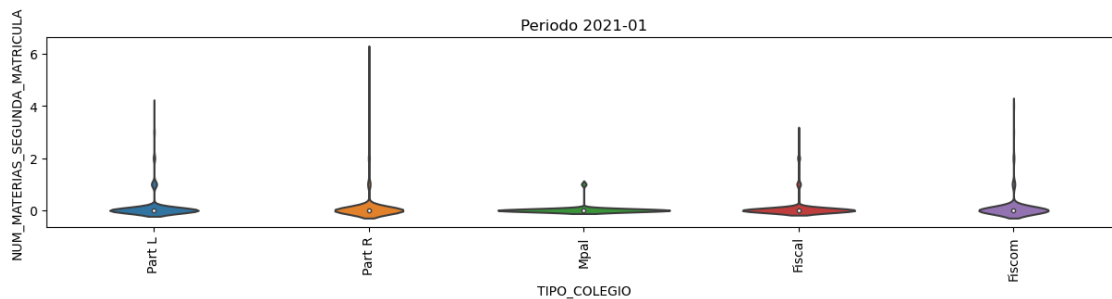


Ilustración 352

Diagrama de Violín del segundo periodo de la relación entre TIPO_COLEGIO y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas

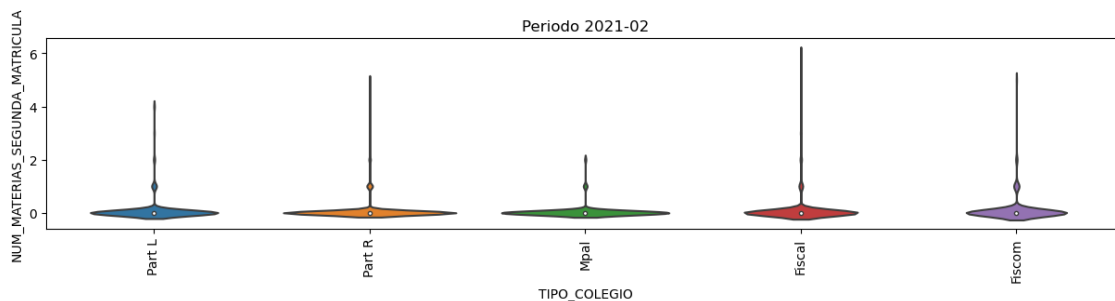


Ilustración 353

Diagrama de Violín del tercer periodo de la relación entre TIPO_COLEGIO y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas

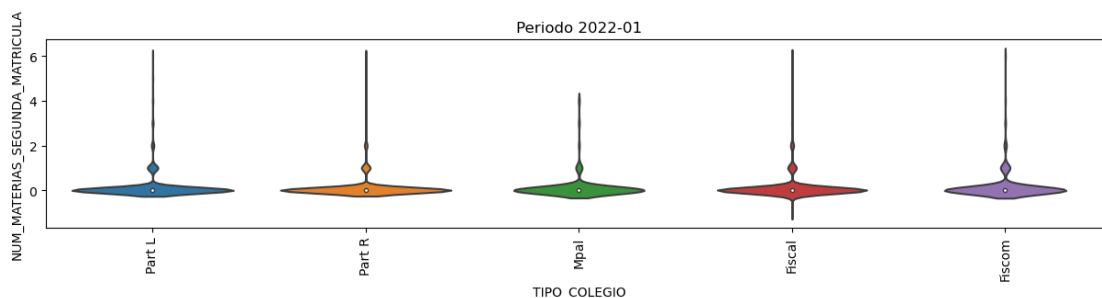
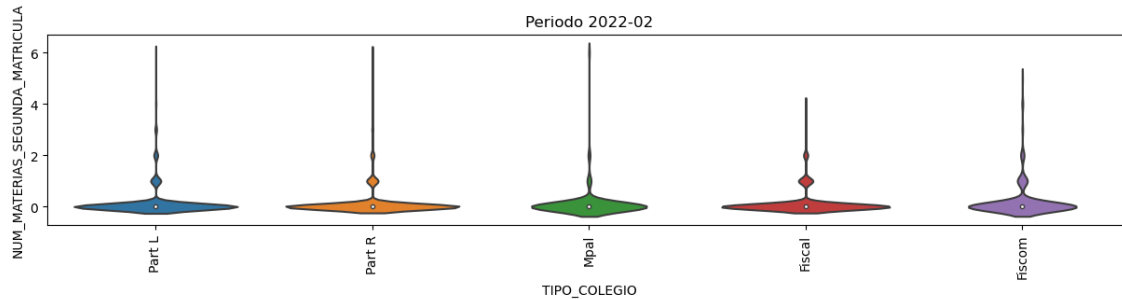


Ilustración 354

Diagrama de Violín del cuarto periodo de la relación entre TIPO_COLEGIO y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas



En conclusión, se puede observar claramente que no existe una aparente relación entre el tipo de colegio al que perteneció un estudiante determinado y la cantidad de asignaturas con segunda matrícula que posee. En contraste, la evaluación de la relación entre el nivel académico del estudiante y su cantidad de asignaturas con segunda matrícula dio como resultado lo siguiente:

Ilustración 355

Código que genera los gráficos para la relación entre NIVEL_POR_MATERIAS y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas

```
# Grid para Los gráficos
fig, axes = plt.subplots(4, 1, figsize=(15, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='NIVEL_POR_MATERIAS', y='NUM_MATERIAS_SEGUNDA_MATRICULA', data=becas[0], ax=axes[0])
axes[0].set_xticklabels(axes[0].get_xticklabels(), rotation=90)

axes[1].set_title('Periodo 2021-02')
sns.violinplot(x='NIVEL_POR_MATERIAS', y='NUM_MATERIAS_SEGUNDA_MATRICULA', data=becas[1], ax=axes[1])
axes[1].set_xticklabels(axes[1].get_xticklabels(), rotation=90)

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='NIVEL_POR_MATERIAS', y='NUM_MATERIAS_SEGUNDA_MATRICULA', data=becas[2], ax=axes[2])
axes[2].set_xticklabels(axes[2].get_xticklabels(), rotation=90)

axes[3].set_title('Periodo 2022-02')
sns.violinplot(x='NIVEL_POR_MATERIAS', y='NUM_MATERIAS_SEGUNDA_MATRICULA', data=becas[3], ax=axes[3])
axes[3].set_xticklabels(axes[3].get_xticklabels(), rotation=90)

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.8)

plt.show() # Mostrar los gráficos
```

Ilustración 356

Diagrama de Violín del primer periodo de la relación entre NIVEL_POR_MATERIAS y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas

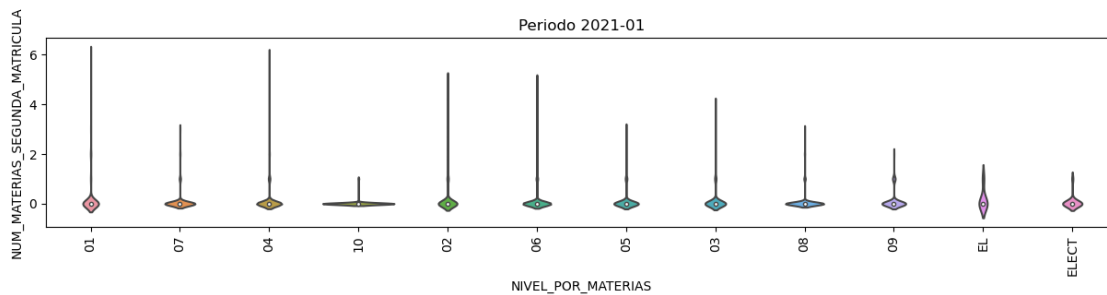


Ilustración 357

Diagrama de Violín del segundo periodo de la relación entre NIVEL_POR_MATERIAS y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas

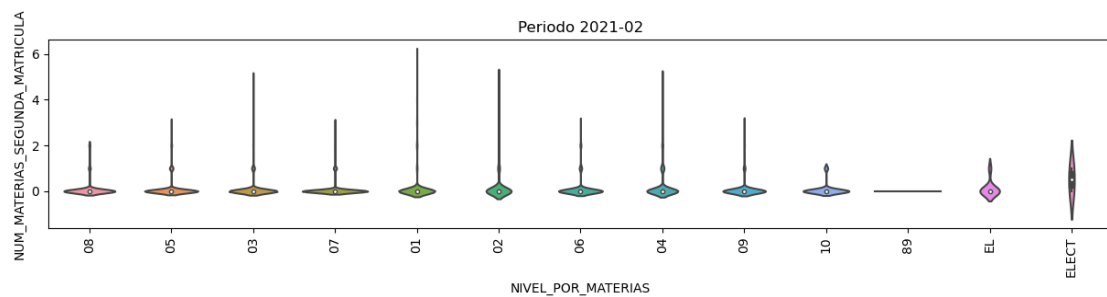


Ilustración 358

Diagrama de Violín del tercer periodo de la relación entre NIVEL_POR_MATERIAS y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas

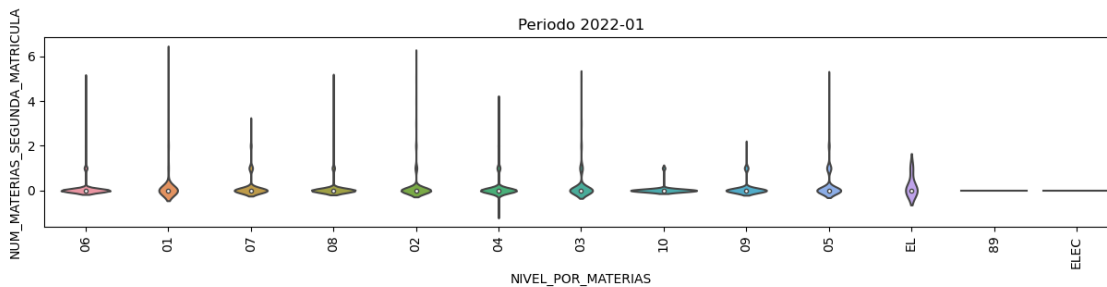
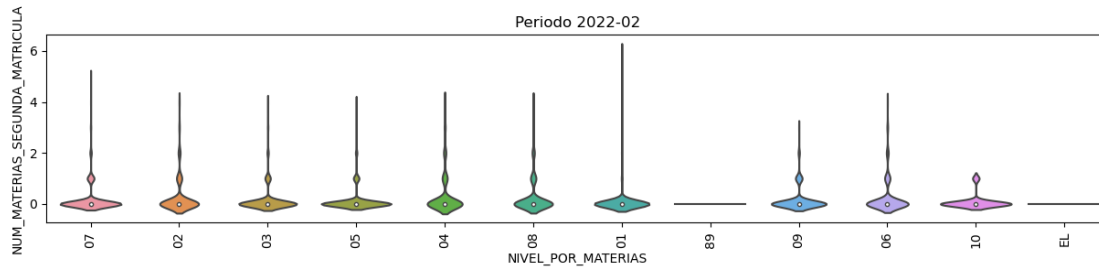


Ilustración 359

Diagrama de Violín del cuarto periodo de la relación entre NIVEL_POR_MATERIAS y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas



En conclusión, se puede observar que en esta ocasión si existe una relación entre el nivel académico de un estudiante y su cantidad de asignaturas con segunda matrícula. De igual manera, la evaluación de la relación entre el nivel de una asignatura cursada por un estudiante determinado y su cantidad de asignaturas con segunda matrícula ofreció el siguiente resultado:

Ilustración 360

Código que genera los gráficos para la relación entre NIVEL_POR_ASIG_APROBADAS y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas

```
# Grid para los gráficos
fig, axes = plt.subplots(4, 1, figsize=(15, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='NIVEL_POR_ASIG_APROBADAS', y='NUM_MATERIAS_SEGUNDA_MATRICULA', data=becas[0], ax=axes[0])
axes[0].set_xticklabels(axes[0].get_xticklabels(), rotation=90)

axes[1].set_title('Periodo 2021-02')
sns.violinplot(x='NIVEL_POR_ASIG_APROBADAS', y='NUM_MATERIAS_SEGUNDA_MATRICULA', data=becas[1], ax=axes[1])
axes[1].set_xticklabels(axes[1].get_xticklabels(), rotation=90)

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='NIVEL_POR_ASIG_APROBADAS', y='NUM_MATERIAS_SEGUNDA_MATRICULA', data=becas[2], ax=axes[2])
axes[2].set_xticklabels(axes[2].get_xticklabels(), rotation=90)

axes[3].set_title('Periodo 2022-02')
sns.violinplot(x='NIVEL_POR_ASIG_APROBADAS', y='NUM_MATERIAS_SEGUNDA_MATRICULA', data=becas[3], ax=axes[3])
axes[3].set_xticklabels(axes[3].get_xticklabels(), rotation=90)

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.8)

plt.show() # Mostrar los gráficos
```


Ilustración 361

Diagrama de Violín del primer periodo de la relación entre NIVEL_POR_ASIG_APROBADAS y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas

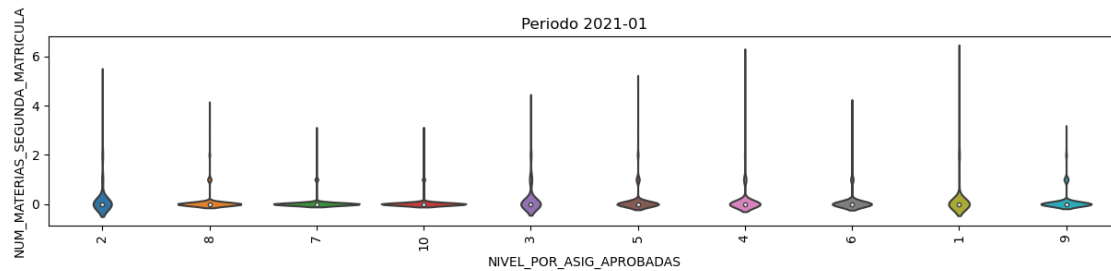


Ilustración 362

Diagrama de Violín del segundo periodo de la relación entre NIVEL_POR_ASIG_APROBADAS y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas

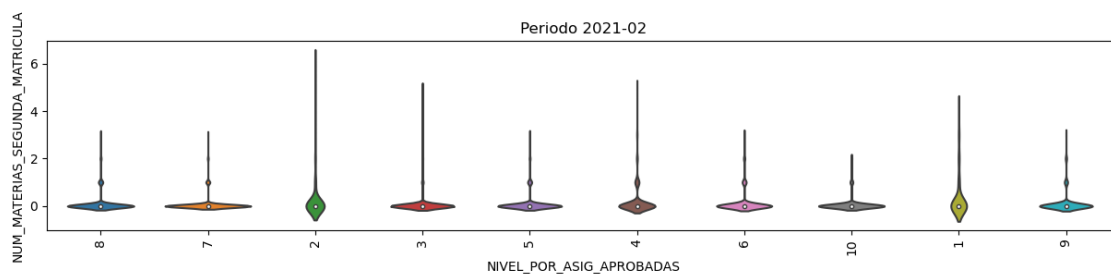


Ilustración 363

Diagrama de Violín del tercer periodo de la relación entre NIVEL_POR_ASIG_APROBADAS y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas

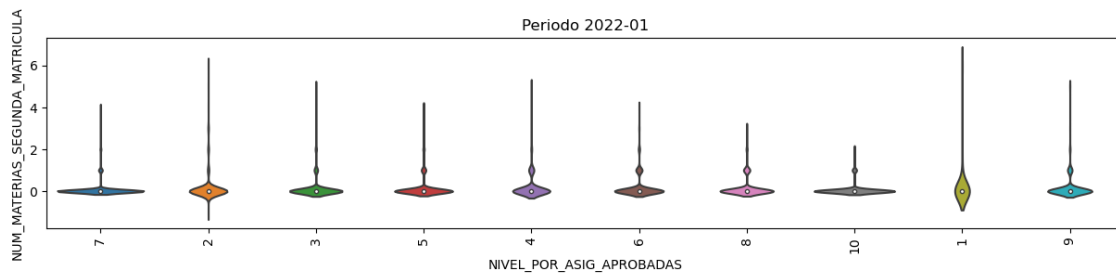
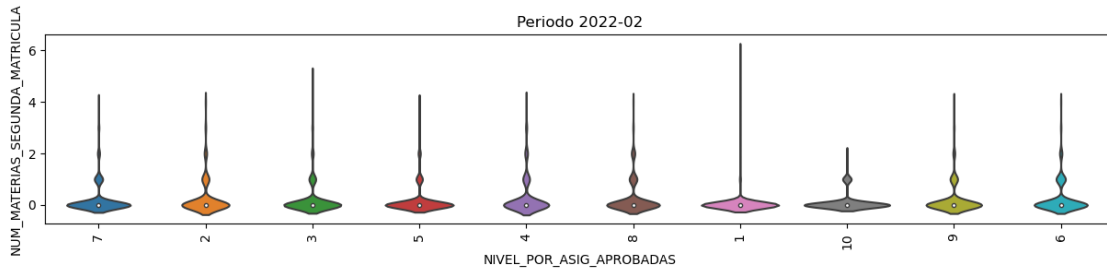


Ilustración 364

Diagrama de Violín del cuarto periodo de la relación entre NIVEL_POR_ASIG_APROBADAS y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas



En conclusión, se puede observar que, de manera semejante al nivel académico de los estudiantes, se presenta una relación entre el nivel de una asignatura cursada por un estudiante determinado y su cantidad de asignaturas con segunda matrícula. Sin embargo, en contraste, la evaluación de la relación entre la posesión de una beca socioeconómica de un estudiante determinado y su cantidad de asignaturas con segunda matrícula dio como resultado lo siguiente:

Ilustración 365

Código que genera los gráficos para la relación entre CATEGORIA_BECA_SOCIOECONOMICA y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas

```
# Grid para los gráficos
fig, axes = plt.subplots(4, 1, figsize=(15, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='CATEGORIA_BECA_SOCIOECONOMICA', y='NUM_MATERIAS_SEGUNDA_MATRICULA', data=becas[0], ax=axes[0])
axes[0].set_xticklabels(axes[0].get_xticklabels(), rotation=90)

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='CATEGORIA_BECA_SOCIOECONOMICA', y='NUM_MATERIAS_SEGUNDA_MATRICULA', data=becas[2], ax=axes[2])
axes[2].set_xticklabels(axes[2].get_xticklabels(), rotation=90)

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.8)

plt.show() # Mostrar los gráficos
```

Ilustración 366

Diagrama de Violín del primer periodo de la relación entre

CATEGORIA_BECA_SOCIOECONOMICA y NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas

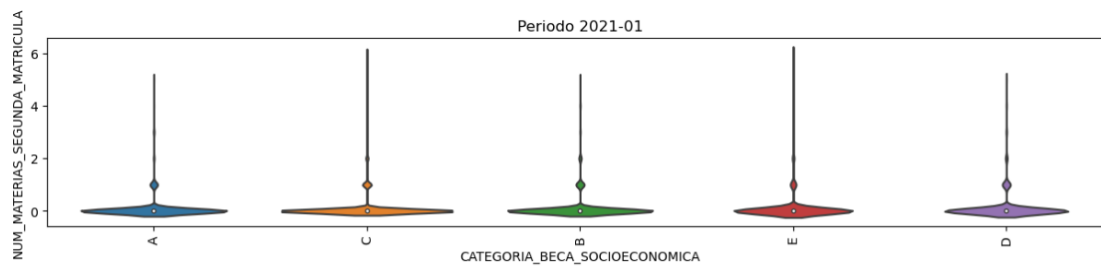
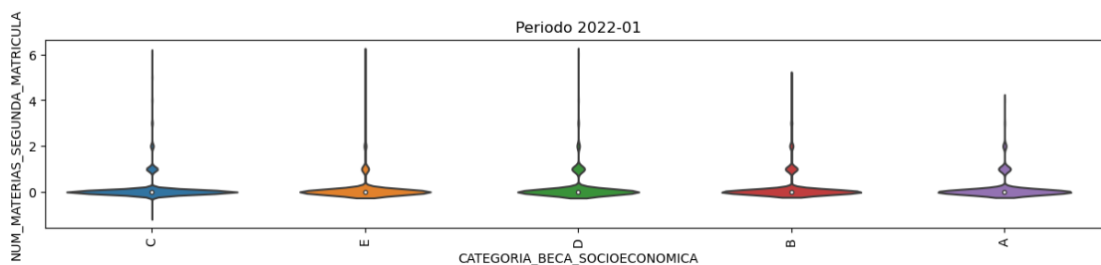


Ilustración 367

Diagrama de Violín del tercer periodo de la relación entre
CATEGORIA_BECA_SOCIOECONOMICA y
NUM_MATERIAS_SEGUNDA_MATRICULA del reporte de Becas



Para empezar, se puede observar que en el código no se presenta la opción de mostrar los gráficos con respecto a los periodos 2021-02 y 2022-02, esto se debe a que como ya se evidencio, estos periodos contienen 0 registros de becas, por otra parte, se puede observar como la cantidad de asignaturas con segunda matrícula no se ve influenciada por las categorías de beca socioeconómica, esto indica que la categoría de beca socioeconómica no parece tener un impacto significativo en la cantidad de asignaturas con segunda matrícula.

Al realizar el análisis de los gráficos por periodo, se puede observar que, en cada uno de los dos, la distribución de la cantidad de asignaturas con segunda matrícula se mantiene prácticamente igual, es decir, no se aprecia una variación significativa en la forma o posición de las distribuciones entre las diferentes categorías de beca.



Ilustración 370

Diagrama de Violín del segundo periodo de la relación entre GENERO y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas

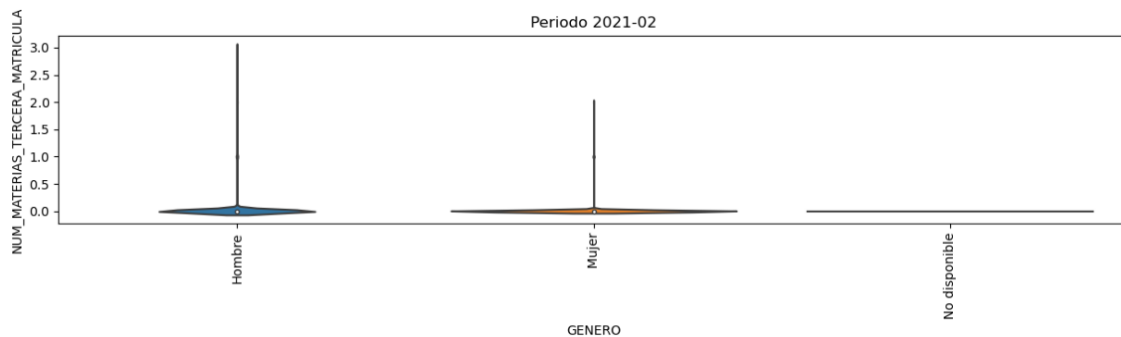


Ilustración 371

Diagrama de Violín del tercer periodo de la relación entre GENERO y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas

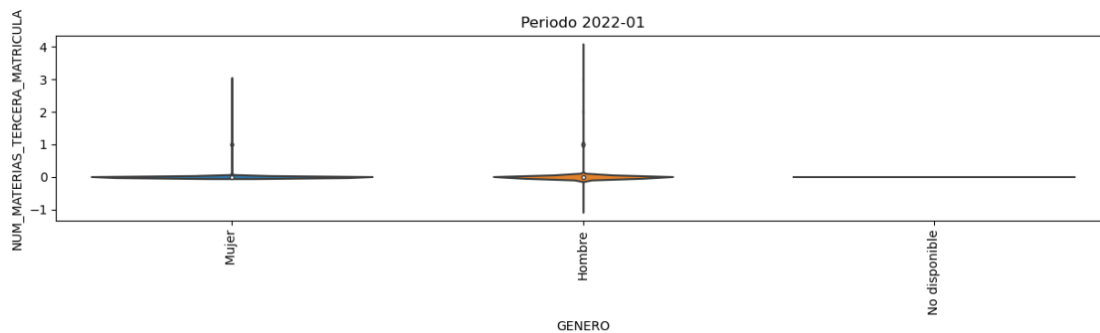
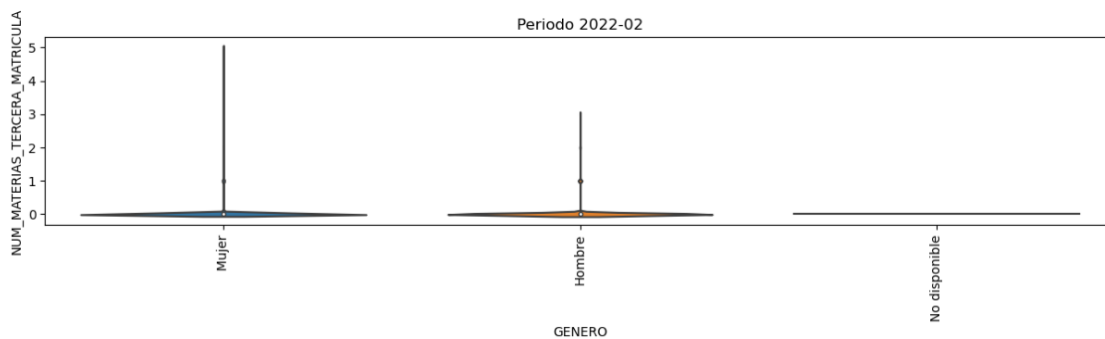


Ilustración 372

Diagrama de Violín del cuarto periodo de la relación entre GENERO y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas



En conclusión, como se puede observar, tampoco existe una relación clara entre el género de un estudiante y su cantidad de asignaturas con tercera matrícula. Por otro lado, con respecto la

evaluación de la relación entre la etnia de un estudiante y su cantidad de asignaturas con tercera matrícula ofreció el siguiente resultado:

Ilustración 373

Código que genera los gráficos para la relación entre ETNIA y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas

```
# Grid para los gráficos
fig, axes = plt.subplots(4, 1, figsize=(15, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='ETNIA', y='NUM_MATERIAS_TERCERA_MATRICULA', data=becas[0], ax=axes[0])
axes[0].set_xticklabels(axes[0].get_xticklabels(), rotation=90)

axes[1].set_title('Periodo 2021-02')
sns.violinplot(x='ETNIA', y='NUM_MATERIAS_TERCERA_MATRICULA', data=becas[1], ax=axes[1])
axes[1].set_xticklabels(axes[1].get_xticklabels(), rotation=90)

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='ETNIA', y='NUM_MATERIAS_TERCERA_MATRICULA', data=becas[2], ax=axes[2])
axes[2].set_xticklabels(axes[2].get_xticklabels(), rotation=90)

axes[3].set_title('Periodo 2022-02')
sns.violinplot(x='ETNIA', y='NUM_MATERIAS_TERCERA_MATRICULA', data=becas[3], ax=axes[3])
axes[3].set_xticklabels(axes[3].get_xticklabels(), rotation=90)

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.8)

plt.show() # Mostrar los gráficos
```

Ilustración 374

Diagrama de Violín del primer periodo de la relación entre ETNIA y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas

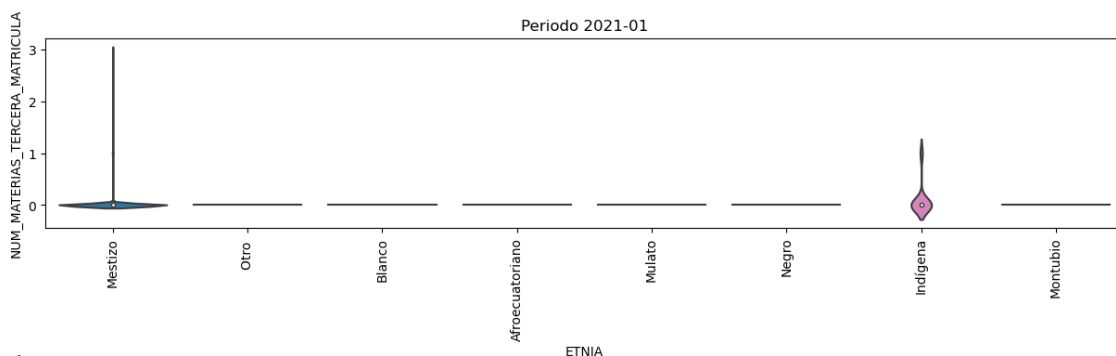


Ilustración 375

Diagrama de Violín del segundo periodo de la relación entre ETNIA y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas

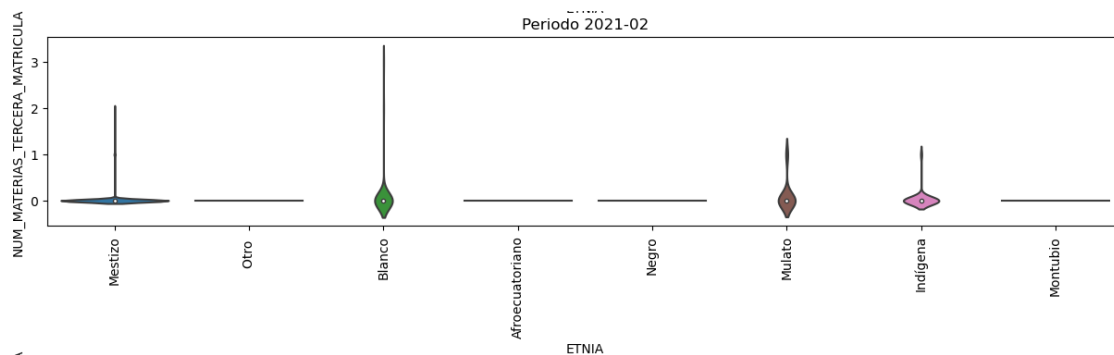


Ilustración 376

Diagrama de Violín del tercer periodo de la relación entre ETNIA y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas

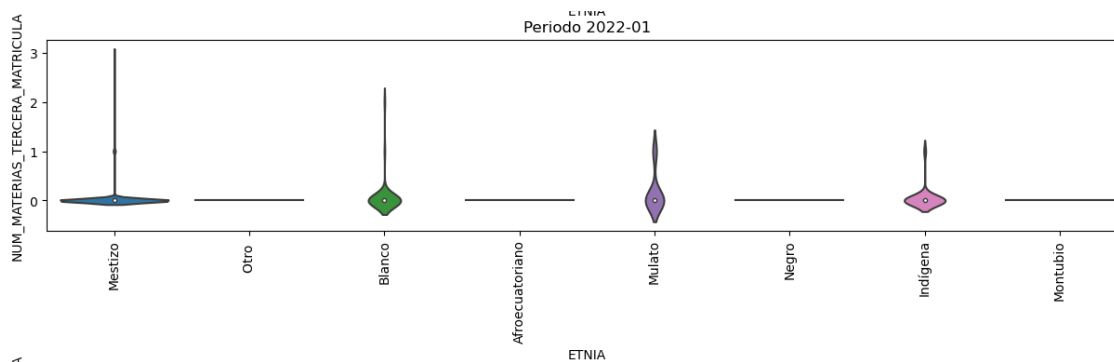
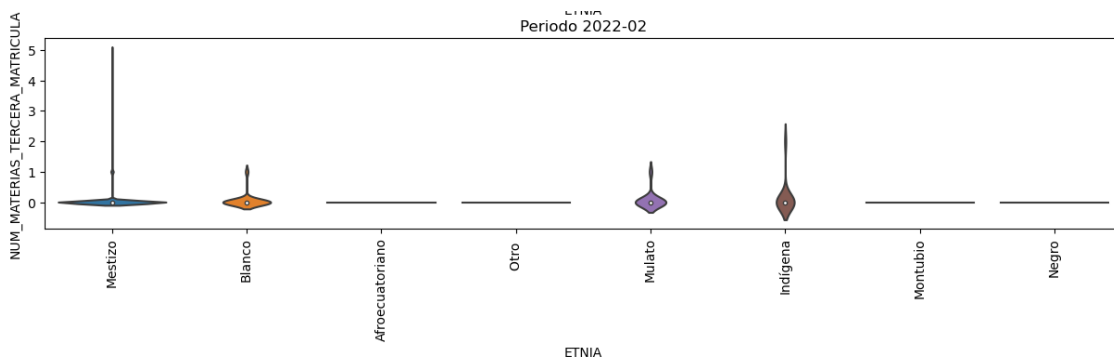


Ilustración 377

Diagrama de Violín del cuarto periodo de la relación entre ETNIA y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas



En conclusión, se puede observar que a pesar de ser menor la distribución de datos con respecto a la cantidad de asignaturas con tercera matrícula que posee un estudiante, los gráficos presentan una relación con respecto a su etnia. Sin embargo, a diferencia de esta evaluación, la

evaluación de la relación entre el tipo de colegio al que asistió un estudiante y su cantidad de asignaturas con tercera matrícula dio el siguiente resultado:

Ilustración 378

Código que genera los gráficos para la relación entre TIPO_COLEGIO y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas

```
# Grid para Los gráficos
fig, axes = plt.subplots(4, 1, figsize=(15, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='TIPO_COLEGIO', y='NUM_MATERIAS_TERCERA_MATRICULA', data=becas[0], ax=axes[0])
axes[0].set_xticklabels(axes[0].get_xticklabels(), rotation=90)

axes[1].set_title('Periodo 2021-02')
sns.violinplot(x='TIPO_COLEGIO', y='NUM_MATERIAS_TERCERA_MATRICULA', data=becas[1], ax=axes[1])
axes[1].set_xticklabels(axes[1].get_xticklabels(), rotation=90)

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='TIPO_COLEGIO', y='NUM_MATERIAS_TERCERA_MATRICULA', data=becas[2], ax=axes[2])
axes[2].set_xticklabels(axes[2].get_xticklabels(), rotation=90)

axes[3].set_title('Periodo 2022-02')
sns.violinplot(x='TIPO_COLEGIO', y='NUM_MATERIAS_TERCERA_MATRICULA', data=becas[3], ax=axes[3])
axes[3].set_xticklabels(axes[3].get_xticklabels(), rotation=90)

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.8)

plt.show() # Mostrar Los gráficos
```

Ilustración 379

Diagrama de Violín del primer periodo de la relación entre TIPO_COLEGIO y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas

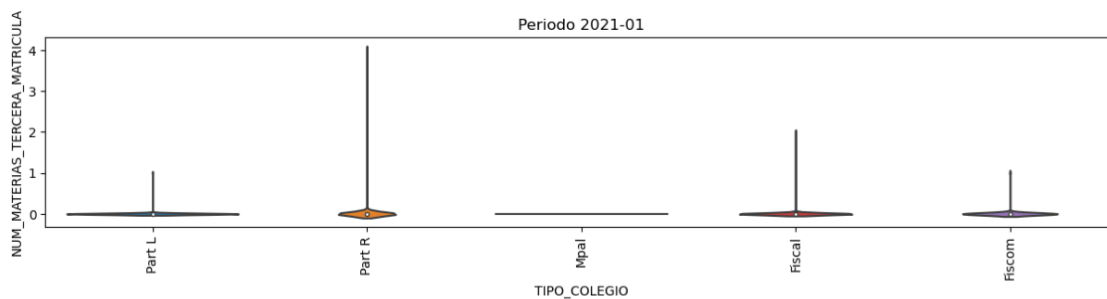


Ilustración 380

Diagrama de Violín del segundo periodo de la relación entre TIPO_COLEGIO y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas

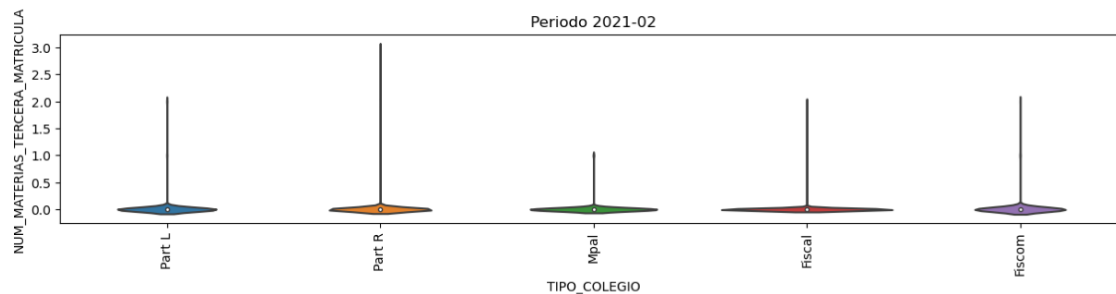


Ilustración 381

Diagrama de Violín del tercer periodo de la relación entre TIPO_COLEGIO y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas

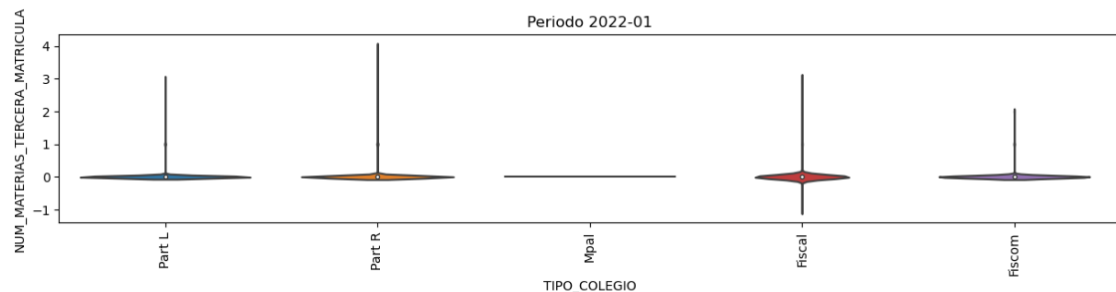
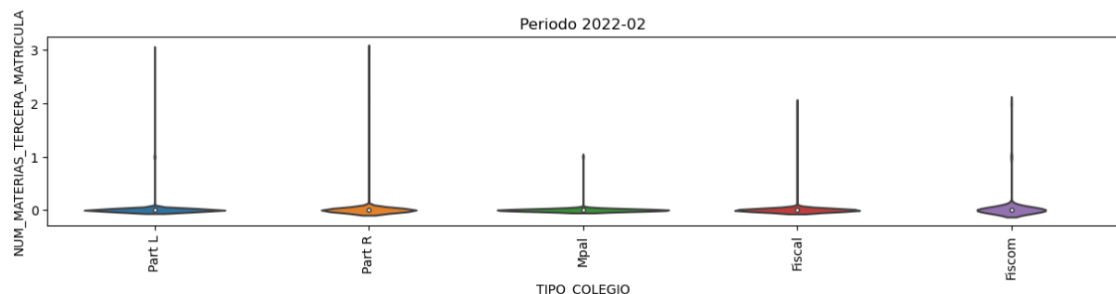


Ilustración 382

Diagrama de Violín del cuarto periodo de la relación entre TIPO_COLEGIO y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas



Se puede observar que no existe una relación entre el colegio donde estudio un estudiante y su cantidad de asignaturas con tercera matrícula. Por otra parte, en contraste, la evaluación de la relación entre el nivel académico del estudiante y su cantidad de asignaturas con tercera matrícula dio el siguiente resultado:

Ilustración 383

Código que genera los gráficos para la relación entre NIVEL_POR_MATERIAS y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas

```
# Grid para Los gráficos
fig, axes = plt.subplots(4, 1, figsize=(15, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='NIVEL_POR_MATERIAS', y='NUM_MATERIAS_TERCERA_MATRICULA', data=becas[0], ax=axes[0])
axes[0].set_xticklabels(axes[0].get_xticklabels(), rotation=90)

axes[1].set_title('Periodo 2021-02')
sns.violinplot(x='NIVEL_POR_MATERIAS', y='NUM_MATERIAS_TERCERA_MATRICULA', data=becas[1], ax=axes[1])
axes[1].set_xticklabels(axes[1].get_xticklabels(), rotation=90)

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='NIVEL_POR_MATERIAS', y='NUM_MATERIAS_TERCERA_MATRICULA', data=becas[2], ax=axes[2])
axes[2].set_xticklabels(axes[2].get_xticklabels(), rotation=90)

axes[3].set_title('Periodo 2022-02')
sns.violinplot(x='NIVEL_POR_MATERIAS', y='NUM_MATERIAS_TERCERA_MATRICULA', data=becas[3], ax=axes[3])
axes[3].set_xticklabels(axes[3].get_xticklabels(), rotation=90)

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.8)

plt.show() # Mostrar Los gráficos
```

Ilustración 384

Diagrama de Violín del primer periodo de la relación entre NIVEL_POR_MATERIAS y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas

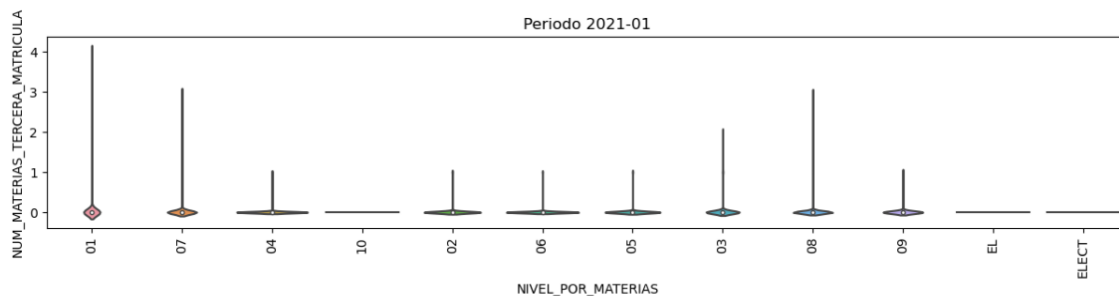


Ilustración 385

Diagrama de Violín del segundo periodo de la relación entre NIVEL_POR_MATERIAS y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas

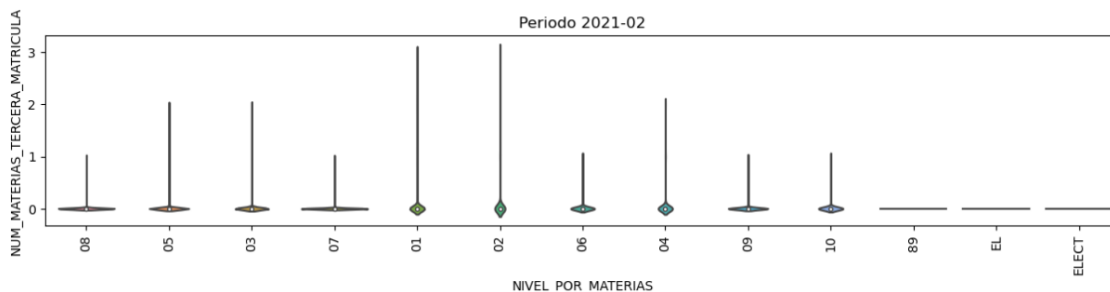


Ilustración 386

Diagrama de Violín del tercer periodo de la relación entre NIVEL_POR_MATERIAS y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas

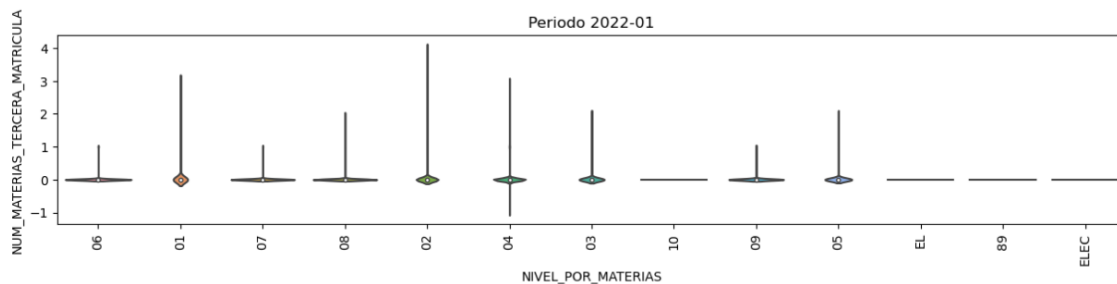
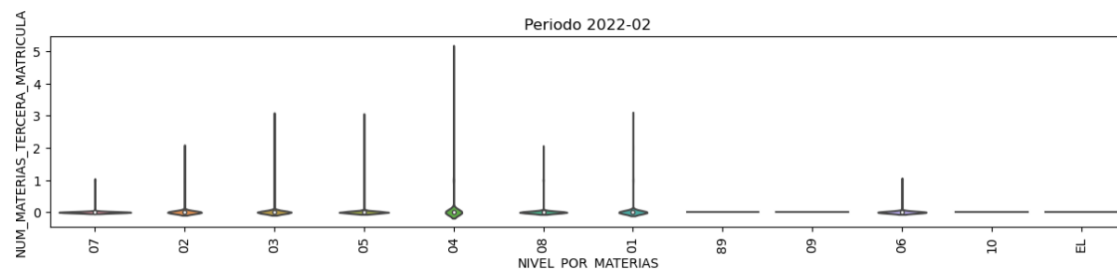


Ilustración 387

Diagrama de Violín del cuarto periodo de la relación entre NIVEL_POR_MATERIAS y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas



En conclusión, se puede observar que existe una posible relación entre el nivel académico de un estudiante y su cantidad de asignaturas con tercera matrícula, este resultado se asemeja al provisto en el análisis con respecto a la cantidad de segundas matrículas, por lo que, de igual manera el resultado de la evaluación de la relación entre el nivel de una asignatura cursada por un estudiante determinado y su cantidad de asignaturas con tercera matrícula fue el siguiente:

Ilustración 388

Código que genera los gráficos para la relación entre NIVEL_POR_ASIG_APROBADAS y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas

```
# Grid para los gráficos
fig, axes = plt.subplots(4, 1, figsize=(15, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='NIVEL_POR_ASIG_APROBADAS', y='NUM_MATERIAS_TERCERA_MATRICULA', data=becas[0], ax=axes[0])
axes[0].set_xticklabels(axes[0].get_xticklabels(), rotation=90)

axes[1].set_title('Periodo 2021-02')
sns.violinplot(x='NIVEL_POR_ASIG_APROBADAS', y='NUM_MATERIAS_TERCERA_MATRICULA', data=becas[1], ax=axes[1])
axes[1].set_xticklabels(axes[1].get_xticklabels(), rotation=90)

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='NIVEL_POR_ASIG_APROBADAS', y='NUM_MATERIAS_TERCERA_MATRICULA', data=becas[2], ax=axes[2])
axes[2].set_xticklabels(axes[2].get_xticklabels(), rotation=90)

axes[3].set_title('Periodo 2022-02')
sns.violinplot(x='NIVEL_POR_ASIG_APROBADAS', y='NUM_MATERIAS_TERCERA_MATRICULA', data=becas[3], ax=axes[3])
axes[3].set_xticklabels(axes[3].get_xticklabels(), rotation=90)

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.8)

plt.show() # Mostrar los gráficos
```

Ilustración 389

Diagrama de Violín del primer periodo de la relación entre NIVEL_POR_ASIG_APROBADAS y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas

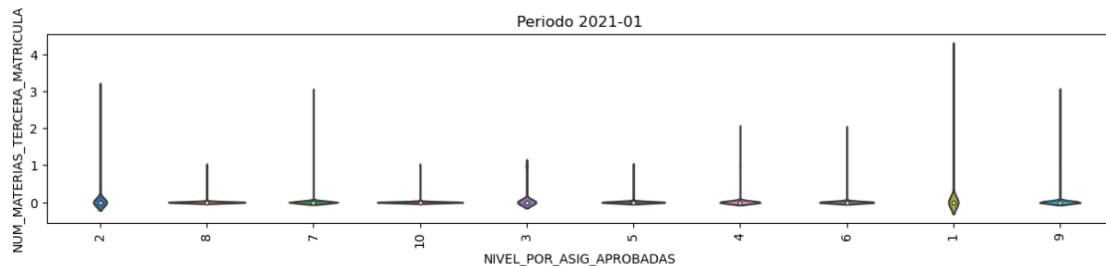


Ilustración 390

Diagrama de Violín del segundo periodo de la relación entre NIVEL_POR_ASIG_APROBADAS y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas

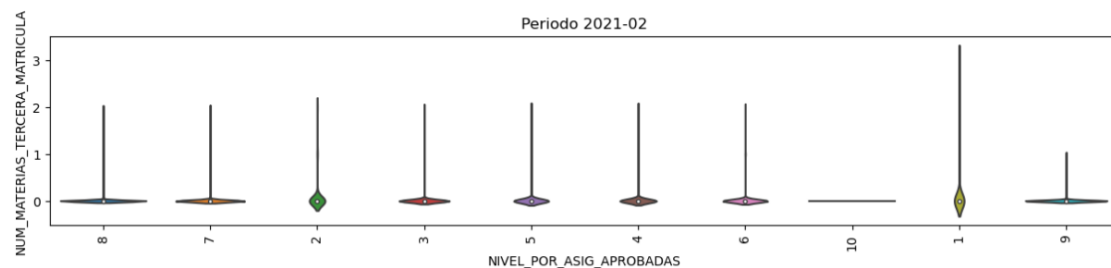


Ilustración 391

Diagrama de Violín del tercer periodo de la relación entre NIVEL_POR_ASIG_APROBADAS y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas

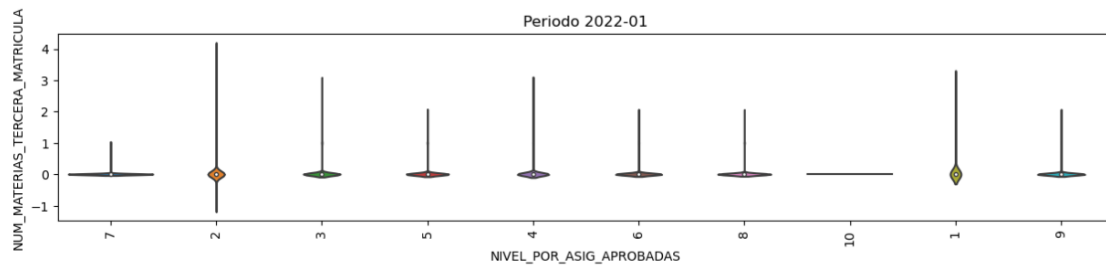
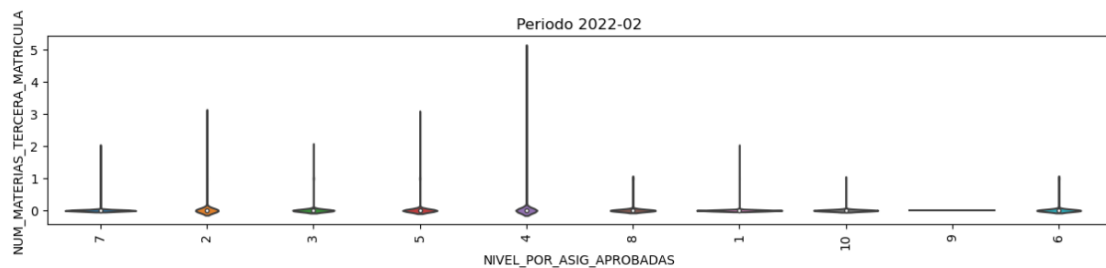


Ilustración 392

Diagrama de Violín del cuarto periodo de la relación entre NIVEL_POR_ASIG_APROBADAS y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas



En conclusión, se puede observar que en los gráficos se presenta una relación entre el nivel de una asignatura aprobada por un estudiante determinado y su cantidad de asignaturas con tercera matrícula, por lo que, se comprueba nuevamente un comportamiento similar a la anterior hipótesis, de hecho, se evidencia igualmente con la evaluación de la relación entre la posesión de una beca socioeconómica de un estudiante determinado y su cantidad de asignaturas con tercera matrícula, que puede ser visualizada a continuación:

Ilustración 393

Código que genera los gráficos para la relación entre

CATEGORIA_BECA_SOCIOECONOMICA y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas

```
# Grid para los gráficos
fig, axes = plt.subplots(4, 1, figsize=(15, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='CATEGORIA_BECA_SOCIOECONOMICA', y='NUM_MATERIAS_TERCERA_MATRICULA', data=becas[0], ax=axes[0])
axes[0].set_xticklabels(axes[0].get_xticklabels(), rotation=90)

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='CATEGORIA_BECA_SOCIOECONOMICA', y='NUM_MATERIAS_TERCERA_MATRICULA', data=becas[2], ax=axes[2])
axes[2].set_xticklabels(axes[2].get_xticklabels(), rotation=90)

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.8)

plt.show() # Mostrar los gráficos
```

Ilustración 394

Diagrama de Violín del primer periodo de la relación entre CATEGORIA_BECA_SOCIOECONOMICA y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas

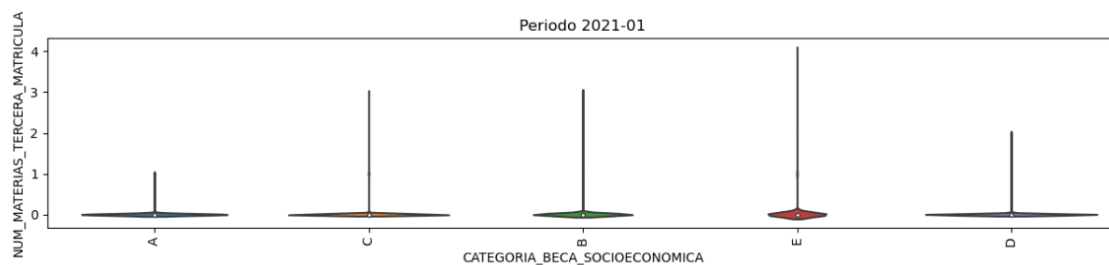
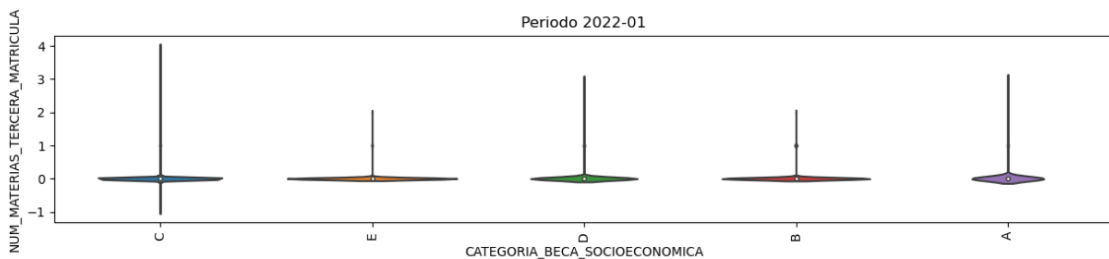


Ilustración 395

Diagrama de Violín del tercer periodo de la relación entre CATEGORIA_BECA_SOCIOECONOMICA y NUM_MATERIAS_TERCERA_MATRICULA del reporte de Becas



Para empezar, se puede observar que en el código nuevamente no se presenta la opción de mostrar los gráficos con respecto a los periodos 2021-02 y 2022-02, esto se debe a que como ya se evidencio, estos periodos contienen 0 registros de becas, por otra parte, se puede observar

cómo la cantidad de asignaturas con tercera matrícula se ve levemente afectada por la categoría de beca socioeconómica que posee el mismo determinado estudiante, sin embargo, al ser la variación realmente mínima entre categoría de beca, se puede argumentar que no existe tal relación.

Finalmente se ha llegado al fin de la exploración de relaciones del reporte de becas, y, para empezar, cabe recalcar que las hipótesis que se debían tratar fueron:

- Las variables referidas al género, etnia, tipo de colegio, carrera, nivel del estudiante, nivel de asignaturas tomadas, categoría de beca socioeconómica influyen en el rendimiento académico con respecto a la cantidad de materias en segunda matrícula.
- Las variables referidas al género, etnia, tipo de colegio, carrera, nivel del estudiante, nivel de asignaturas tomadas, categoría de beca socioeconómica y cantidad de asignaturas con segunda matrícula influyen en el rendimiento académico con respecto a la cantidad de materias en tercera matrícula.

Por lo que, tras los análisis realizados se pudo comprobar que ambas hipótesis tienen un grado de certitud con algunas excepciones, sin embargo, a pesar de los resultados de los análisis realizados hasta el momento, es factible la idea de que las variables referentes al género, etnia, tipo de colegio y beca socioeconómica presenten una influencia en el rendimiento académico en términos diferentes a la cantidad de asignaturas con segunda y tercera matrícula que poseen los estudiantes. Esta teoría se basa principalmente en los resultados del análisis llevado a cabo en el reporte de asignaturas perdidas donde se comprobó que no existe relación entre el rendimiento académico y la cantidad de veces que un estudiante repite una determinada asignatura.

Es por ello por lo que se ha tomado la decisión de realizar un análisis de las relaciones entre las variables del reporte de becas y la calificación final que han tenido estos estudiantes, lo que significa que se llevará a cabo una unión de datos enfocada en evidenciar las calificaciones finales que los estudiantes han obtenido en el reporte de becas. No obstante, se debe mencionar que se obviarán los análisis de relación sobre las características `PROGAMA_DESC`, `NIVEL_ESTUDIANTE` y `AREA_MATERIAS` debido a que repetidamente se ha verificado la existencia de una relación de estas variables con el rendimiento académico.

Por otro lado, con respecto a la unión, esta se realizará entre los reportes de calificaciones por carrera y el de becas, a razón de ello, para llevar a cabo esta unión se utilizarán las variables identificadoras del estudiante y sus notas finales; sin embargo, es importante tener en cuenta que los registros de calificaciones en el reporte de calificaciones por carrera corresponden a todas las asignaturas que un estudiante determinado tomó en el periodo académico, mientras que los registros en el reporte de becas corresponden a un único registro por estudiante. Dada esta situación, para solucionarla, se calculará el promedio de las calificaciones finales del estudiante en todas sus asignaturas cursadas durante el periodo evaluado, donde, posteriormente, este valor promedio será utilizado para realizar la unión con el reporte de becas.

A partir de lo mencionado, se comenzará el proceso de unión con una verificación actual de las dimensiones de los periodos de becas:

Ilustración 396

Verificación actual de dimensiones de los periodos de becas

```
for beca in becas:
    print(f"El reporte {beca.name} tiene {beca.shape[0]} registros y {beca.shape[1]} variables.")

El reporte becas01 tiene 9399 registros y 14 variables.
El reporte becas02 tiene 8564 registros y 14 variables.
El reporte becas03 tiene 8336 registros y 14 variables.
El reporte becas04 tiene 8255 registros y 14 variables.
```


Se puede observar la cantidad de registros y variables que posee cada uno de los marcos de datos del reporte de becas, a partir de ello, se procederá a realizar la unión en marcos de datos temporales, teniendo en cuenta que únicamente se utilizarán las variables `id_banner_estudiante` y `nota_final` del reporte de calificaciones por carrera.

Ilustración 397

Realización de la unión en marcos de datos temporales

```
# Almacenará los dataframes unidos
becas_unidas = []

# Se realiza una iteración sobre todos los periodos de los reportes
for i in range(4):
    # Creamos un dataframe temporal con únicamente las variables requeridas
    temporal_df = calificaciones[i][["id_banner_estudiante", "nota_final"]].copy()

    # Cambiamos los nombres de las variables del dataframe temporal
    temporal_df.columns = ["ID_BANNER", "NOTA_FINAL"]

    # Calculamos la nota final promedio de cada estudiante que se encuentre en ambos reportes
    promedios = []
    estudiantes = becas[i]["ID_BANNER"].unique()
    for estudiante in estudiantes:
        promedio = round(temporal_df["NOTA_FINAL"][temporal_df["ID_BANNER"] == estudiante].mean(), 2)
        promedios.append(promedio)

    # Se crea un nuevo dataframe con los datos de los estudiantes y sus calificaciones finales promedios
    datos_df = pd.DataFrame({"ID_BANNER":estudiantes, "NOTA_FINAL":promedios})

    # Unimos los conjuntos de datos
    temp = pd.merge(datos_df, becas[i], on="ID_BANNER", how="inner")
    becas_unidas.append(temp)
```

Una vez se crearon los marcos de datos temporales, se consideró realizar una verificación en código que cerciore que la unión fue un éxito, la misma puede ser visualizada en la primera ilustración siguiente; sin embargo, se consideró de igual forma realizar una verificación visual de los datos, por lo que, después de la primera ilustración que sigue a este párrafo, usted estimado lector, podrá observar y verificar por su cuenta un marco de datos temporal con respecto al original.

Ilustración 398

Verificación de los marcos de datos temporales con respecto a los marcos de datos originales

```
for i in range(4):
    print(f"El reporte becas{becas[i].name} es igual al reporte nuevo?", becas[i].shape[0] == becas_unidas[i].shape[0])

El reporte becasbecas01 es igual al reporte nuevo? True
El reporte becasbecas02 es igual al reporte nuevo? True
El reporte becasbecas03 es igual al reporte nuevo? True
El reporte becasbecas04 es igual al reporte nuevo? True
```

Ilustración 399

Parte 1 de la vista del marco de datos del primer periodo del reporte de Becas original

becas[0].head()									
	ID_BANNER	NACIONALIDAD	GENERO	ETNIA	TIPO_COLEGIO	PROGAMA_DESC	AREA_MATERIAS	NIVEL_POR_MATERIAS	NIVEL_POR_ASIG_APROBADAS
0	E9532	Ecuatoriano	Mujer	NaN	NaN	Enfermería	01	01	2
1	E2371	Ecuatoriano	Hombre	Mestizo	NaN	Economía	07	07	8
2	E2838	Ecuatoriano	Mujer	NaN	Part L	Negocios Internacionales	04	04	7
3	E3781	Ecuatoriano	Mujer	Mestizo	NaN	Arquitectura	10	10	10
4	E2249	Ecuatoriano	Hombre	NaN	NaN	Teología	02	02	2

Ilustración 400

Parte final de la vista del marco de datos del primer periodo del reporte de Becas original

CATEGORIA_BECA_SOCIOECONOMICA	NUM_MATERIAS_SEGUNDA_MATRICULA	NUM_MATERIAS_TERCERA_MATRICULA	TIENE_PAGO	TIENE_SALDO	
A	1	0	SI	NO	
C	0	0	SI	NO	
C	1	0	SI	NO	
A	0	0	SI	NO	
B	0	0	SI	NO	

Ilustración 401

Parte 1 de la vista del marco de datos del primer periodo del reporte de Becas temporal

becas_unidas[0].head()										
	ID_BANNER	NOTA_FINAL	NACIONALIDAD	GENERO	ETNIA	TIPO_COLEGIO	PROGAMA_DESC	AREA_MATERIAS	NIVEL_POR_MATERIAS	NIVEL_POR_ASIG_APROBADAS
0	E9532	29.53	Ecuatoriano	Mujer	NaN	NaN	Enfermería	01	01	2
1	E2371	41.57	Ecuatoriano	Hombre	Mestizo	NaN	Economía	07	07	8
2	E2838	44.98	Ecuatoriano	Mujer	NaN	Part L	Negocios Internacionales	04	04	7
3	E3781	45.93	Ecuatoriano	Mujer	Mestizo	NaN	Arquitectura	10	10	10
4	E2249	43.19	Ecuatoriano	Hombre	NaN	NaN	Teología	02	02	2

Ilustración 402

Parte final de la vista del marco de datos del primer periodo del reporte de Becas temporal

CATEGORIA_BECA_SOCIOECONOMICA	NUM_MATERIAS_SEGUNDA_MATRICULA	NUM_MATERIAS_TERCERA_MATRICULA	TIENE_PAGO	TIENE_SALDO
A	1	0	SI	NO
C	0	0	SI	NO
C	1	0	SI	NO
A	0	0	SI	NO
B	0	0	SI	NO

Finalmente, dado que ambas verificaciones establecieron que no existen problemas evidentes, se llevó concreto la unión en los marcos de datos oficiales.

Ilustración 403

Integración de la unión en los datos oficiales

```
for i in range(4):  
    becas[i] = becas_unidas[i]  
    becas[i].name = "becas0" + str(i)
```

Ilustración 404

Parte 1 de la vista del marco de datos unificado del primer periodo del reporte de Becas

becas[0].head()										
	ID_BANNER	NOTA_FINAL	NACIONALIDAD	GENERO	ETNIA	TIPO_COLEGIO	PROGAMA_DESC	AREA_MATERIAS	NIVEL_POR_MATERIAS	NIVEL_POR_ASIG_APROBADAS
0	E9532	29.53	Ecuatoriano	Mujer	NaN	NaN	Enfermería	01	01	2
1	E2371	41.57	Ecuatoriano	Hombre	Mestizo	NaN	Economía	07	07	8
2	E2838	44.98	Ecuatoriano	Mujer	NaN	Part L	Negocios Internacionales	04	04	7
3	E3781	45.93	Ecuatoriano	Mujer	Mestizo	NaN	Arquitectura	10	10	10
4	E2249	43.19	Ecuatoriano	Hombre	NaN	NaN	Teología	02	02	2

Ilustración 405

Parte final de la vista del marco de datos unificado del primer periodo del reporte de Becas

CATEGORIA_BECA_SOCIOECONOMICA	NUM_MATERIAS_SEGUNDA_MATRICULA	NUM_MATERIAS_TERCERA_MATRICULA	TIENE_PAGO	TIENE_SALDO
A	1	0	SI	NO
C	0	0	SI	NO
C	1	0	SI	NO
A	0	0	SI	NO
B	0	0	SI	NO

Ahora, una vez se concretó la unificación planteada, es momento de plantear una nueva hipótesis con la nueva variable disponible:

- Las variables referidas al género, etnia, tipo de colegio, nivel del estudiante, nivel de asignaturas tomadas, categoría de beca socioeconómica, cantidad de asignaturas con segunda matrícula y cantidad de asignaturas con tercera matrícula influyen en el rendimiento académico en términos de la calificación final que han tenido estos estudiantes.

Por lo que, para empezar, se evaluó la relación entre el género de un estudiante y su calificación final, cuyo resultado es el siguiente:

Ilustración 406

Código que genera los gráficos para la relación entre GENERO y NOTA_FINAL del reporte de Becas

```
# Grid para los gráficos
fig, axes = plt.subplots(4, 1, figsize=(15, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='GENERO', y='NOTA_FINAL', order=["Mujer", "Hombre", "No disponible"], data=becas[0], ax=axes[0])
axes[0].set_xticklabels(axes[0].get_xticklabels(), rotation=90)

axes[1].set_title('Periodo 2021-02')
sns.violinplot(x='GENERO', y='NOTA_FINAL', order=["Mujer", "Hombre", "No disponible"], data=becas[1], ax=axes[1])
axes[1].set_xticklabels(axes[1].get_xticklabels(), rotation=90)

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='GENERO', y='NOTA_FINAL', order=["Mujer", "Hombre", "No disponible"], data=becas[2], ax=axes[2])
axes[2].set_xticklabels(axes[2].get_xticklabels(), rotation=90)

axes[3].set_title('Periodo 2022-02')
sns.violinplot(x='GENERO', y='NOTA_FINAL', order=["Mujer", "Hombre", "No disponible"], data=becas[3], ax=axes[3])
axes[3].set_xticklabels(axes[3].get_xticklabels(), rotation=90)

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.8)

plt.show() # Mostrar los gráficos
```

Ilustración 407

Diagrama de Violín del primer periodo de la relación entre GENERO y NOTA_FINAL del reporte de Becas

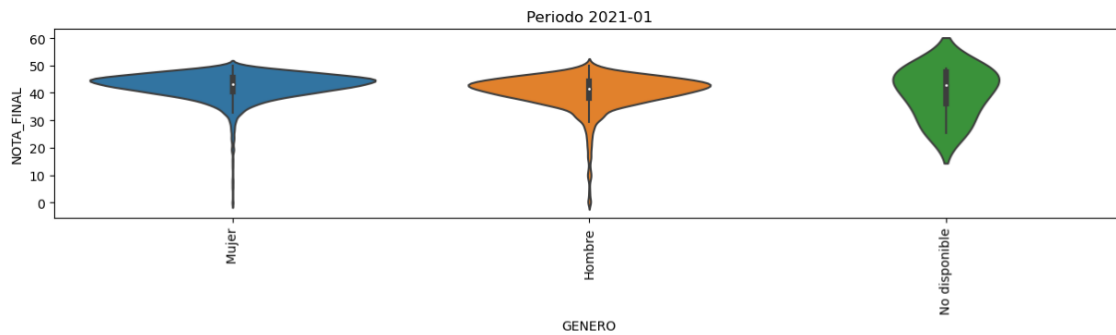


Ilustración 408

Diagrama de Violín del segundo periodo de la relación entre GENERO y NOTA_FINAL del reporte de Becas

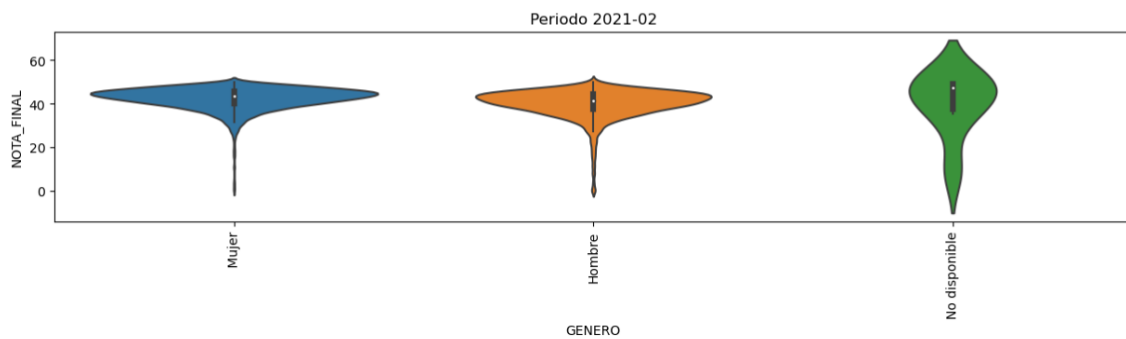


Ilustración 409

Diagrama de Violín del tercer periodo de la relación entre GENERO y NOTA_FINAL del reporte de Becas

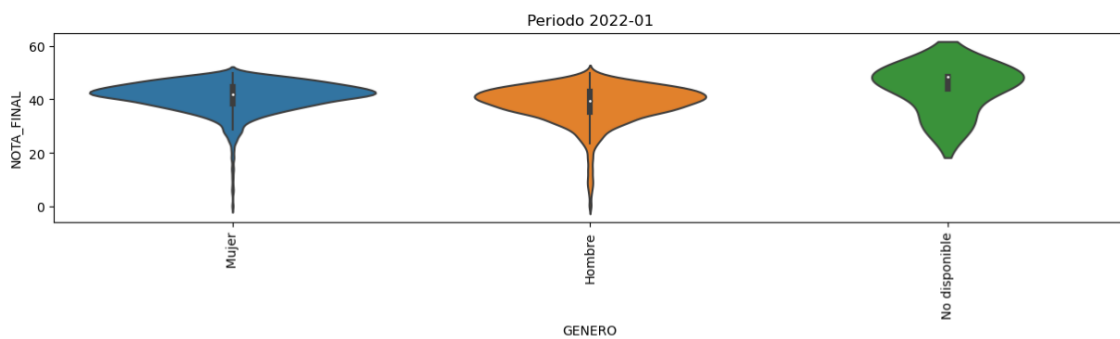
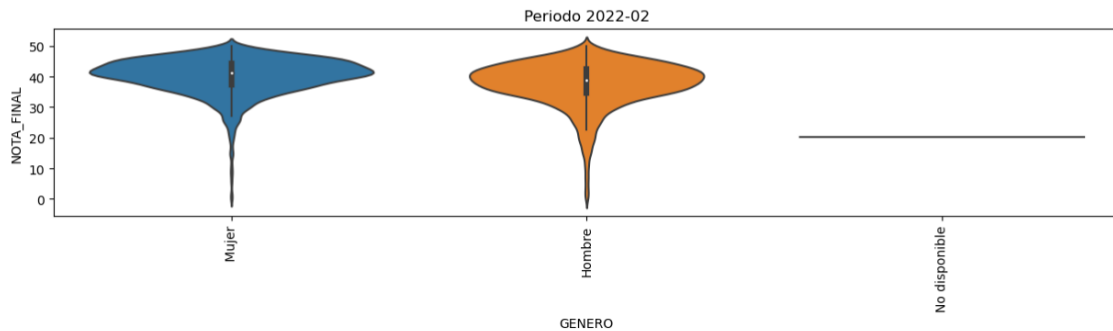


Ilustración 410

Diagrama de Violín del cuarto periodo de la relación entre GENERO y NOTA_FINAL del reporte de Becas



En conclusión, se puede concretar finalmente que no existe una relación entre el rendimiento académico y el género de un estudiante, ya que las diferencias que presentan los patrones de figura de ambos géneros únicamente varían de acuerdo con la cantidad de estudiantes correspondientes a cada uno. Por otro lado, la evaluación de la relación entre la etnia de un estudiante y sus calificaciones finales dio como resultado el siguiente:

Ilustración 411

Código que genera los gráficos para la relación entre ETNIA y NOTA_FINAL del reporte de Becas

```
# Grid para Los gráficos
fig, axes = plt.subplots(4, 1, figsize=(15, 17))

# Generamos un orden para la muestra de Los gráficos
orden = becas[0]["ETNIA"].unique().tolist()
orden.pop(0)

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='ETNIA', y='NOTA_FINAL', order=orden, data=becas[0], ax=axes[0])
axes[0].set_xticklabels(axes[0].get_xticklabels(), rotation=90)

axes[1].set_title('Periodo 2021-02')
sns.violinplot(x='ETNIA', y='NOTA_FINAL', order=orden, data=becas[1], ax=axes[1])
axes[1].set_xticklabels(axes[1].get_xticklabels(), rotation=90)

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='ETNIA', y='NOTA_FINAL', order=orden, data=becas[2], ax=axes[2])
axes[2].set_xticklabels(axes[2].get_xticklabels(), rotation=90)

axes[3].set_title('Periodo 2022-02')
sns.violinplot(x='ETNIA', y='NOTA_FINAL', order=orden, data=becas[3], ax=axes[3])
axes[3].set_xticklabels(axes[3].get_xticklabels(), rotation=90)

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.8)

plt.show() # Mostrar Los gráficos
```

Ilustración 412

Diagrama de Violín del primer periodo de la relación entre ETNIA y NOTA_FINAL del reporte de Becas

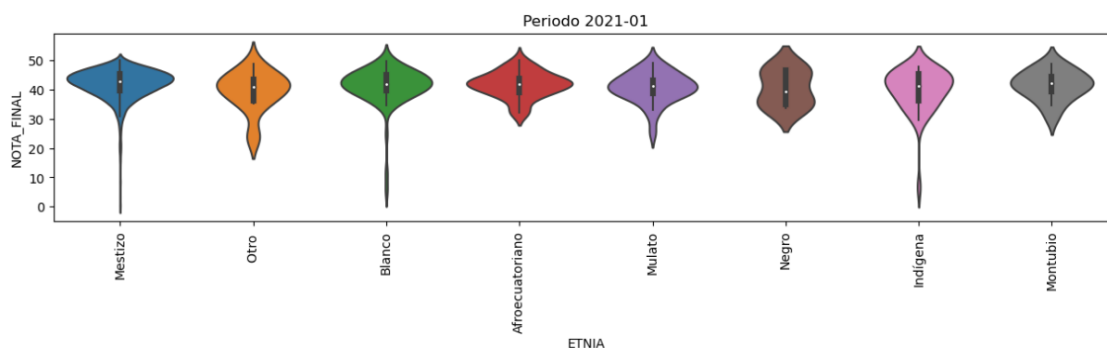


Ilustración 413

Diagrama de Violín del segundo periodo de la relación entre ETNIA y NOTA_FINAL del reporte de Becas

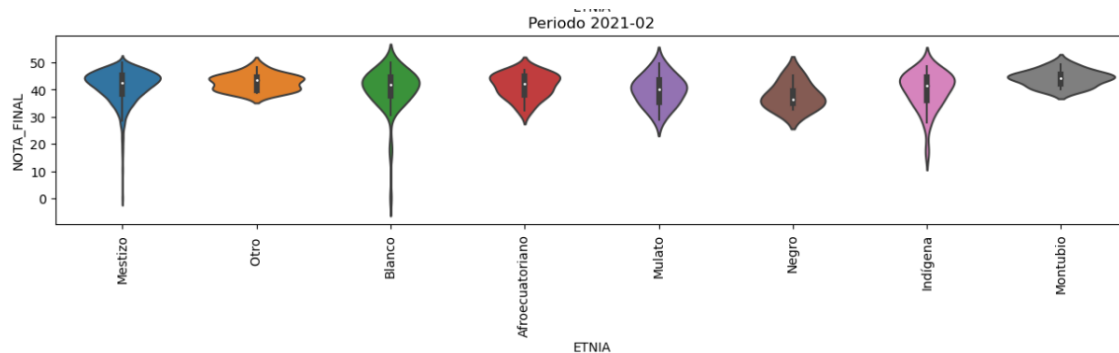


Ilustración 414

Diagrama de Violín del tercer periodo de la relación entre ETNIA y NOTA_FINAL del reporte de Becas

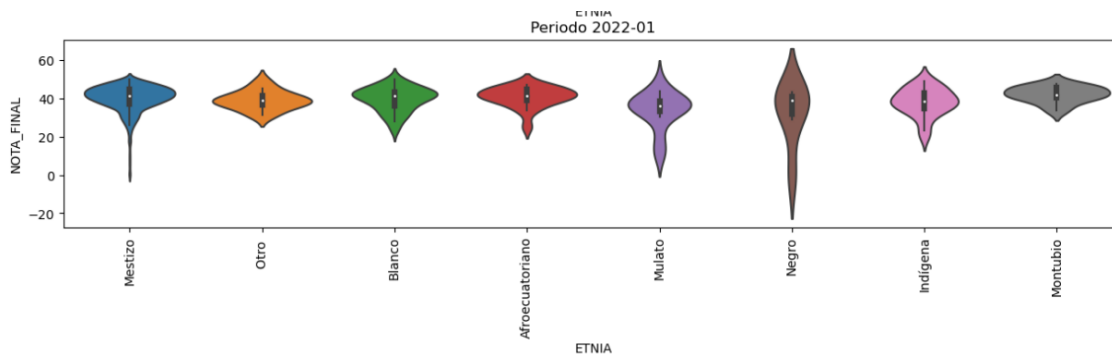
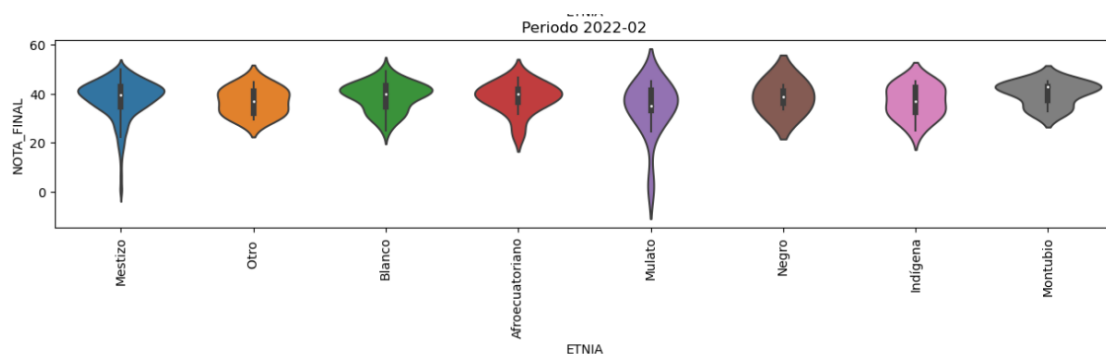


Ilustración 415

Diagrama de Violín del cuarto periodo de la relación entre ETNIA y NOTA_FINAL del reporte de Becas



Una vez más, es posible identificar una ligera asociación entre la etnia y el rendimiento académico, no obstante, es importante destacar que esta correlación no parece ser

particularmente fuerte, dado que la variabilidad en ciertos casos es apenas perceptible. A pesar de esto, la relación es innegable y es más notoria que en sus anteriores análisis.

Sin embargo, no se debería descartar la hipótesis de que, tras la mínima discrepancia observada en los patrones de figura de las distribuciones, podría sugerir que la etnia no es necesariamente un factor significativo al determinar el rendimiento académico de un estudiante, no obstante, para esta investigación, será considerada como una variable relevante dado los resultados de los análisis llevados a cabo sobre ella.

Por otro lado, la relación entre el tipo de colegio al que asistió un estudiante y sus calificaciones finales ofreció el siguiente resultado:

Ilustración 416

Código que genera los gráficos para la relación entre TIPO_COLEGIO y NOTA_FINAL del reporte de Becas

```
# Grid para los gráficos
fig, axes = plt.subplots(4, 1, figsize=(15, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='TIPO_COLEGIO', y='NOTA_FINAL', data=becas[0], ax=axes[0])
axes[0].set_xticklabels(axes[0].get_xticklabels(), rotation=90)

axes[1].set_title('Periodo 2021-02')
sns.violinplot(x='TIPO_COLEGIO', y='NOTA_FINAL', data=becas[1], ax=axes[1])
axes[1].set_xticklabels(axes[1].get_xticklabels(), rotation=90)

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='TIPO_COLEGIO', y='NOTA_FINAL', data=becas[2], ax=axes[2])
axes[2].set_xticklabels(axes[2].get_xticklabels(), rotation=90)

axes[3].set_title('Periodo 2022-02')
sns.violinplot(x='TIPO_COLEGIO', y='NOTA_FINAL', data=becas[3], ax=axes[3])
axes[3].set_xticklabels(axes[3].get_xticklabels(), rotation=90)

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.8)

plt.show() # Mostrar los gráficos
```

Ilustración 417

Diagrama de Violín del primer periodo de la relación entre TIPO_COLEGIO y NOTA_FINAL del reporte de Becas

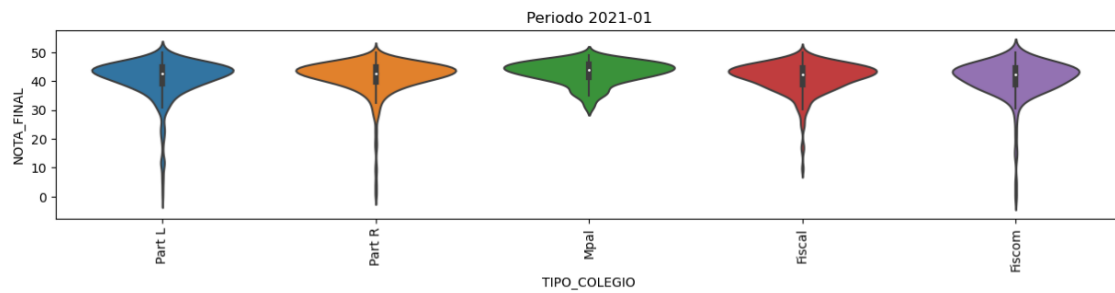


Ilustración 418

Diagrama de Violín del segundo periodo de la relación entre TIPO_COLEGIO y NOTA_FINAL del reporte de Becas

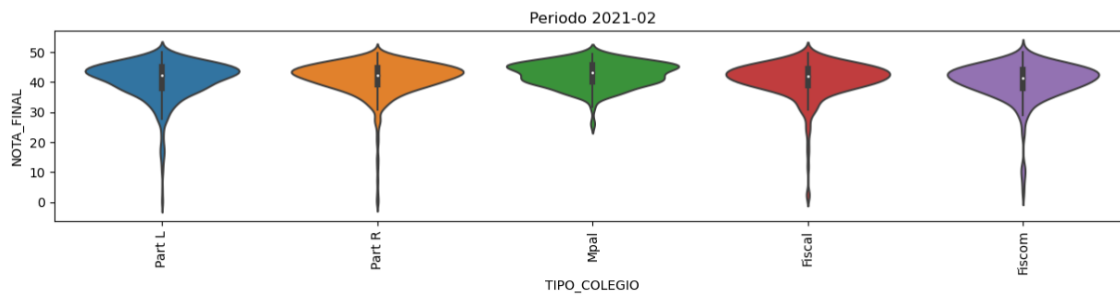


Ilustración 419

Diagrama de Violín del tercer periodo de la relación entre TIPO_COLEGIO y NOTA_FINAL del reporte de Becas

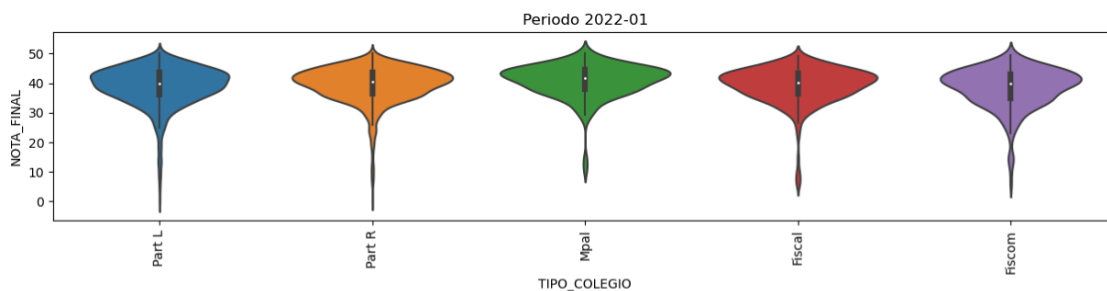
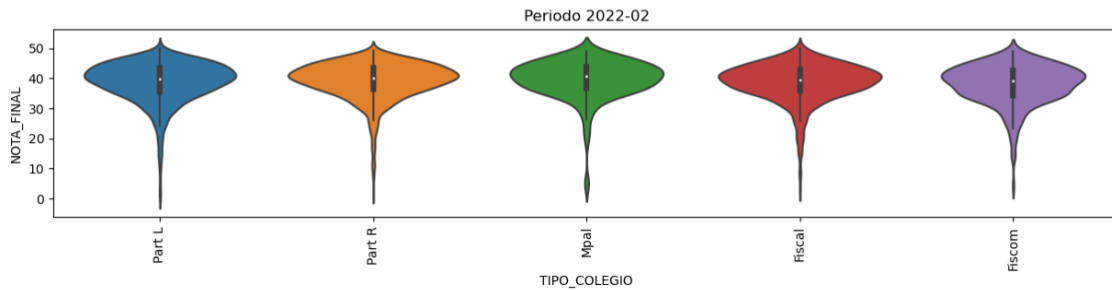


Ilustración 420

Diagrama de Violín del cuarto periodo de la relación entre TIPO_COLEGIO y NOTA_FINAL del reporte de Becas



En conclusión, se puede observar nuevamente que no existe una relación entre el tipo de colegio en el que estuvo un estudiante determinado y su rendimiento académico.

Adicionalmente, la evaluación de la relación entre el nivel de una asignatura aprobada por un estudiante determinado y sus calificaciones finales demostró lo siguiente:

Ilustración 421

Código que genera los gráficos para la relación entre NIVEL_POR_ASIG_APROBADAS y NOTA_FINAL del reporte de Becas

```
# Grid para Los gráficos
fig, axes = plt.subplots(4, 1, figsize=(15, 17))

# Generamos un orden para la muestra de Los gráficos
orden = becas[0]["NIVEL_POR_ASIG_APROBADAS"].unique().tolist()

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='NIVEL_POR_ASIG_APROBADAS', y='NOTA_FINAL', order=orden, data=becas[0], ax=axes[0])
axes[0].set_xticklabels(axes[0].get_xticklabels(), rotation=90)

axes[1].set_title('Periodo 2021-02')
sns.violinplot(x='NIVEL_POR_ASIG_APROBADAS', y='NOTA_FINAL', order=orden, data=becas[1], ax=axes[1])
axes[1].set_xticklabels(axes[1].get_xticklabels(), rotation=90)

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='NIVEL_POR_ASIG_APROBADAS', y='NOTA_FINAL', order=orden, data=becas[2], ax=axes[2])
axes[2].set_xticklabels(axes[2].get_xticklabels(), rotation=90)

axes[3].set_title('Periodo 2022-02')
sns.violinplot(x='NIVEL_POR_ASIG_APROBADAS', y='NOTA_FINAL', order=orden, data=becas[3], ax=axes[3])
axes[3].set_xticklabels(axes[3].get_xticklabels(), rotation=90)

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.8)

plt.show() # Mostrar Los gráficos
```

Ilustración 422

Diagrama de Violín del primer periodo de la relación entre NIVEL_POR_ASIG_APROBADAS y NOTA_FINAL del reporte de Becas

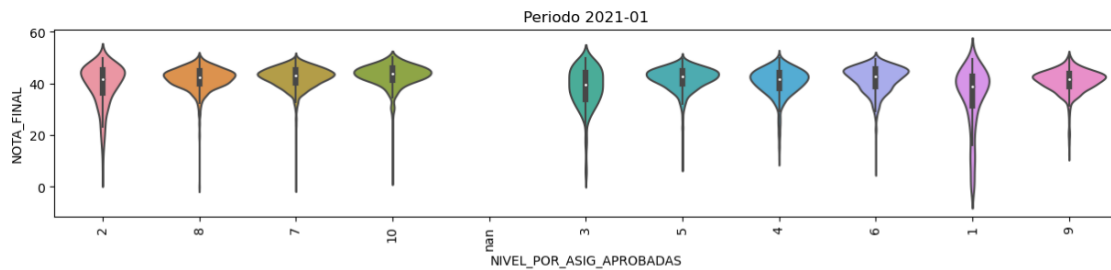


Ilustración 423

Diagrama de Violín del segundo periodo de la relación entre NIVEL_POR_ASIG_APROBADAS y NOTA_FINAL del reporte de Becas

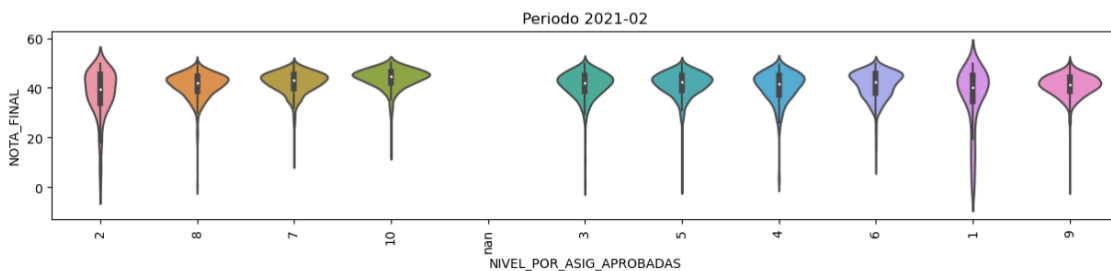


Ilustración 424

Diagrama de Violín del tercer periodo de la relación entre NIVEL_POR_ASIG_APROBADAS y NOTA_FINAL del reporte de Becas

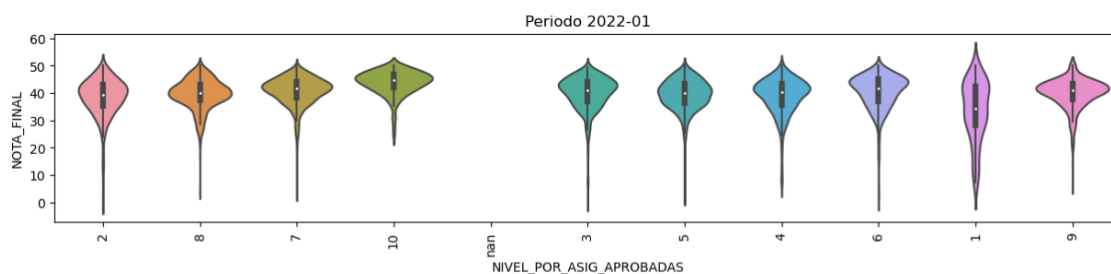
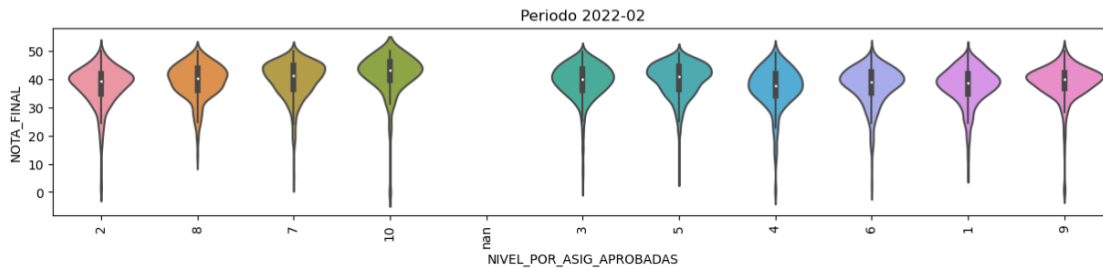


Ilustración 425

Diagrama de Violín del cuarto periodo de la relación entre NIVEL_POR_ASIG_APROBADAS y NOTA_FINAL del reporte de Becas



En conclusión, a pesar de las observaciones dadas en el pasado análisis, se puede determinar que no existe una relación real entre el rendimiento académico y el nivel de una asignatura aprobada. De manera similar, la evaluación de la relación entre la posesión de una beca socioeconómica de un estudiante determinado y sus calificaciones finales exhibió lo siguiente:

Ilustración 426

Código que genera los gráficos para la relación entre CATEGORIA_BECA_SOCIOECONOMICA y NOTA_FINAL del reporte de Becas

```
# Grid para los gráficos
fig, axes = plt.subplots(4, 1, figsize=(15, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='CATEGORIA_BECA_SOCIOECONOMICA', y='NOTA_FINAL', data=becas[0], ax=axes[0])
axes[0].set_xticklabels(axes[0].get_xticklabels(), rotation=90)

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='CATEGORIA_BECA_SOCIOECONOMICA', y='NOTA_FINAL', data=becas[2], ax=axes[2])
axes[2].set_xticklabels(axes[2].get_xticklabels(), rotation=90)

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.8)

plt.show() # Mostrar los gráficos
```

Ilustración 427

Diagrama de Violín del primer periodo de la relación entre CATEGORIA_BECA_SOCIOECONOMICA y NOTA_FINAL del reporte de Becas

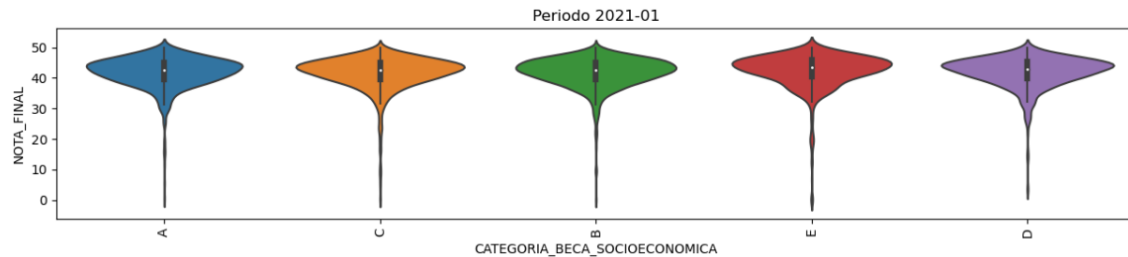
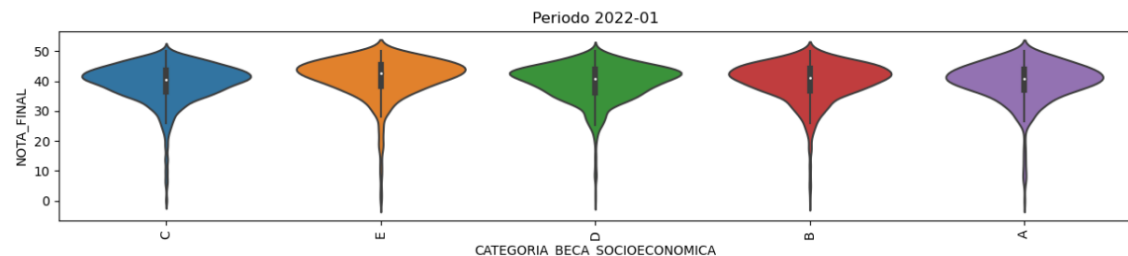


Ilustración 428

Diagrama de Violín del tercer periodo de la relación entre CATEGORIA_BECA_SOCIOECONOMICA y NOTA_FINAL del reporte de Becas



En conclusión, nuevamente es observable que no existe una relación entre la posesión de una beca socioeconómica de un estudiante determinado y su rendimiento académico. En contraste, la evaluación llevada a cabo sobre la relación entre la cantidad de asignaturas con segunda matrícula que posee un estudiante determinado y sus calificaciones finales resultó en lo siguiente:

Ilustración 429

Código que genera los gráficos para la relación entre NUM_MATERIAS_SEGUNDA_MATRICULA y NOTA_FINAL del reporte de Becas

```
# Grid para Los gráficos
fig, axes = plt.subplots(4, 1, figsize=(15, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='NUM_MATERIAS_SEGUNDA_MATRICULA', y='NOTA_FINAL', data=becas[0], ax=axes[0])
axes[0].set_xticklabels(axes[0].get_xticklabels(), rotation=90)

axes[1].set_title('Periodo 2021-02')
sns.violinplot(x='NUM_MATERIAS_SEGUNDA_MATRICULA', y='NOTA_FINAL', data=becas[1], ax=axes[1])
axes[1].set_xticklabels(axes[1].get_xticklabels(), rotation=90)

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='NUM_MATERIAS_SEGUNDA_MATRICULA', y='NOTA_FINAL', data=becas[2], ax=axes[2])
axes[2].set_xticklabels(axes[2].get_xticklabels(), rotation=90)

axes[3].set_title('Periodo 2022-02')
sns.violinplot(x='NUM_MATERIAS_SEGUNDA_MATRICULA', y='NOTA_FINAL', data=becas[3], ax=axes[3])
axes[3].set_xticklabels(axes[3].get_xticklabels(), rotation=90)

# Ajustar el espacio entre los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.8)

plt.show() # Mostrar Los gráficos
```

Ilustración 430

Diagrama de Violín del primer periodo de la relación entre NUM_MATERIAS_SEGUNDA_MATRICULA y NOTA_FINAL del reporte de Becas

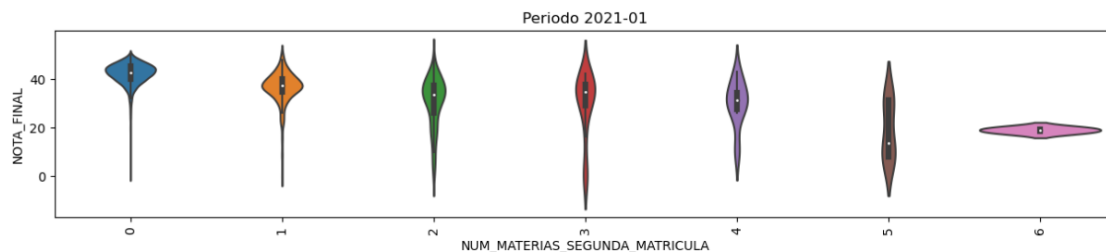


Ilustración 431

Diagrama de Violín del segundo periodo de la relación entre NUM_MATERIAS_SEGUNDA_MATRICULA y NOTA_FINAL del reporte de Becas

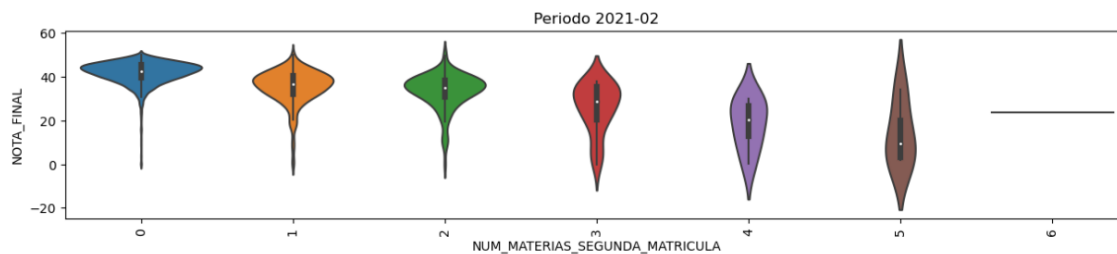


Ilustración 432

Diagrama de Violín del tercer periodo de la relación entre NUM_MATERIAS_SEGUNDA_MATRICULA y NOTA_FINAL del reporte de Becas

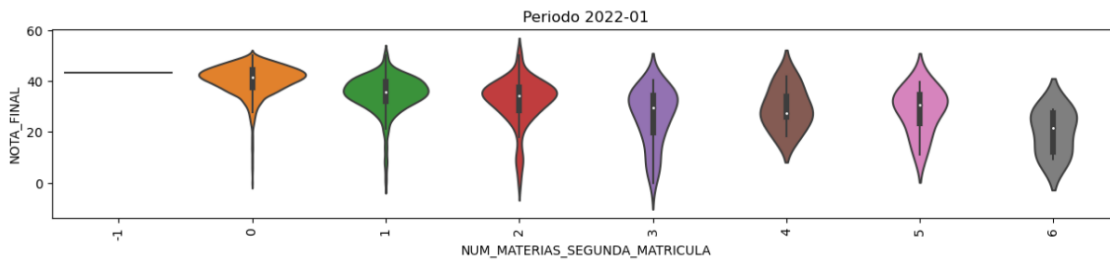
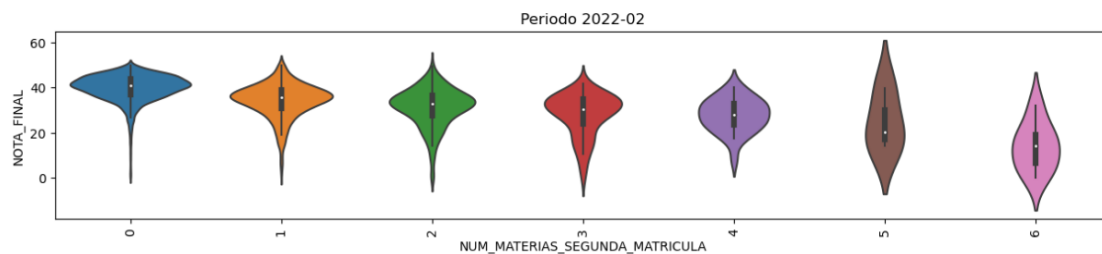


Ilustración 433

Diagrama de Violín del cuarto periodo de la relación entre NUM_MATERIAS_SEGUNDA_MATRICULA y NOTA_FINAL del reporte de Becas



En conclusión, es evidente la existencia de una relación entre el rendimiento académico en términos de calificación final y la cantidad de asignaturas con segunda matrícula, esto podría deberse a afectaciones psicológicas debido a la pérdida de muchas asignaturas. Por lo que, la evaluación de la relación entre la cantidad de asignaturas con tercera matrícula que posee un estudiante determinado y sus calificaciones finales presenta un resultado similar:

Ilustración 434

Código que genera los gráficos para la relación entre
NUM_MATERIAS_TERCERA_MATRICULA y NOTA_FINAL del reporte de Becas

```
# Grid para Los gráficos
fig, axes = plt.subplots(4, 1, figsize=(15, 17))

# Gráficos por periodo
axes[0].set_title('Periodo 2021-01')
sns.violinplot(x='NUM_MATERIAS_TERCERA_MATRICULA', y='NOTA_FINAL', data=becas[0], ax=axes[0])
axes[0].set_xticklabels(axes[0].get_xticklabels(), rotation=90)

axes[1].set_title('Periodo 2021-02')
sns.violinplot(x='NUM_MATERIAS_TERCERA_MATRICULA', y='NOTA_FINAL', data=becas[1], ax=axes[1])
axes[1].set_xticklabels(axes[1].get_xticklabels(), rotation=90)

axes[2].set_title('Periodo 2022-01')
sns.violinplot(x='NUM_MATERIAS_TERCERA_MATRICULA', y='NOTA_FINAL', data=becas[2], ax=axes[2])
axes[2].set_xticklabels(axes[2].get_xticklabels(), rotation=90)

axes[3].set_title('Periodo 2022-02')
sns.violinplot(x='NUM_MATERIAS_TERCERA_MATRICULA', y='NOTA_FINAL', data=becas[3], ax=axes[3])
axes[3].set_xticklabels(axes[3].get_xticklabels(), rotation=90)

# Ajustar el espacio entre Los subplots
plt.subplots_adjust(wspace=0.2, hspace=0.8)

plt.show() # Mostrar Los gráficos
```

Ilustración 435

Diagrama de Violín del primer periodo de la relación entre
NUM_MATERIAS_TERCERA_MATRICULA y NOTA_FINAL del reporte de Becas

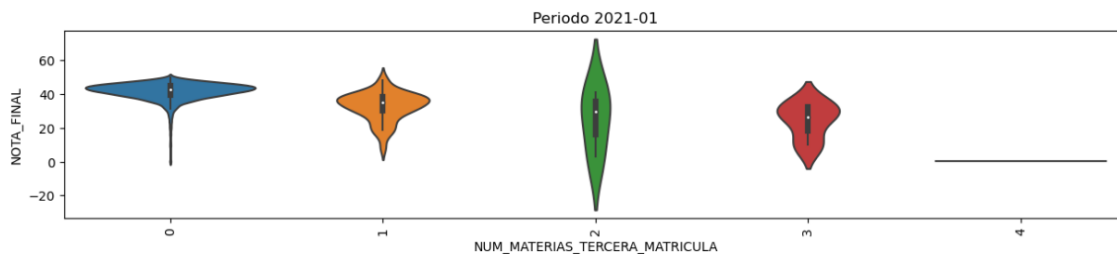


Ilustración 436

Diagrama de Violín del segundo periodo de la relación entre
NUM_MATERIAS_TERCERA_MATRICULA y NOTA_FINAL del reporte de Becas

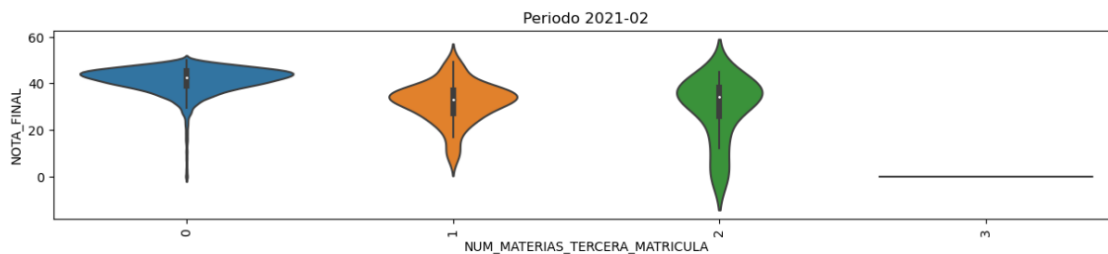


Ilustración 437

Diagrama de Violín del tercer periodo de la relación entre NUM_MATERIAS_TERCERA_MATRICULA y NOTA_FINAL del reporte de Becas

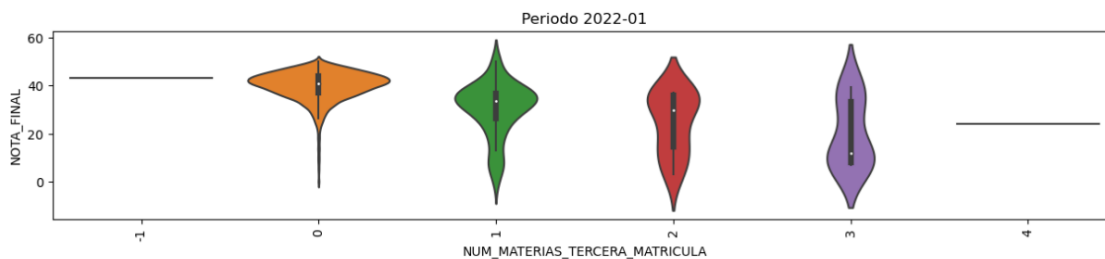
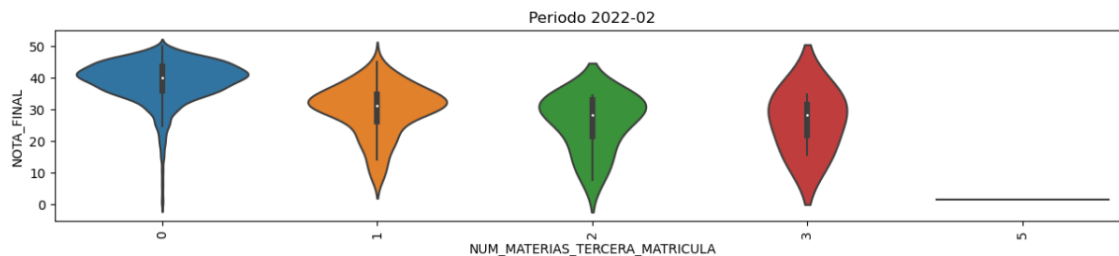


Ilustración 438

Diagrama de Violín del cuarto periodo de la relación entre NUM_MATERIAS_TERCERA_MATRICULA y NOTA_FINAL del reporte de Becas



En conclusión, es evidente la existencia de una relación entre el rendimiento académico en términos de calificación final y la cantidad de asignaturas con tercera matrícula, esto podría deberse a afectaciones psicológicas debido a la pérdida de muchas asignaturas, al igual que en el anterior caso.

Finalmente, se ha llegado al fin de la evaluación de la exploración de las relaciones en el reporte de becas, se debe tener en cuenta que la hipótesis considerada al fin de cuentas fue la siguiente:

- Las variables referidas al género, etnia, tipo de colegio, nivel del estudiante, nivel de asignaturas tomadas, categoría de beca socioeconómica, cantidad de asignaturas con segunda matrícula y cantidad de asignaturas con tercera matrícula influyen en el

rendimiento académico en términos de la calificación final que han tenido estos estudiantes.

Por lo que, tras los análisis realizados se pudo comprobar la hipótesis con algunas excepciones que se pueden observar en la siguiente tabla de relevancia e irrelevancia de las variables:

Tabla 20
Determinación de Irrelevancia y Relevancia de Variables del reporte de Becas

Variable	Relevancia	Comentarios
ID_BANNER	No aplica	Esta variable es ideada para permitir una futura integración de datos.
NOTA_FINAL	No aplica	Esta variable tiene el objetivo de permitir el análisis de las variables que más afectan al rendimiento académico, pero, dado el enfoque de la investigación no resulta útil.
NACIONALIDAD	No relevante	Esta variable fue determinada como no relevante en el análisis de distribución del mismo reporte.
GENERO	No relevante	
ETNIA	Relevante	
TIPO_COLEGIO	No relevante	
NIVEL_POR_MATERIAS	Relevante	
NIVEL_POR_ASIG_APROBADAS	No relevante	
CATEGORIA_BECA_SOCIOECONOMICA	No Relevante	
NUM_MATERIAS_SEGUNDA_MATRICULA	Relevante	
NUM_MATERIAS_TERCERA_MATRICULA	Relevante	

TIENE_PAGO	No relevante	Esta variable fue determinada como no relevante en el análisis de distribución del mismo reporte.
TIENE_SALDO	No relevante	Esta variable fue determinada como no relevante en el análisis de distribución del mismo reporte.
AREA_MATERIAS	No aplica	Esta variable es relevante, sin embargo, se puede obviar su aparición en este reporte dado que los demás ya poseen este registro
PROGAMA_DESC	No aplica	Esta variable es relevante, sin embargo, se puede obviar su aparición en este reporte dado que los demás ya poseen este registro

4.1.13. Selección avanzada de variables relevantes

En el presente subcapítulo, se abordará una selección avanzada de variables relevantes, basados en los resultados de los anteriores capítulos donde se realizaron exploraciones exhaustivas de variables y se examinaron las relaciones, a partir de ello, se procederá a identificar y descartar las variables irrelevantes, este proceso permitirá reducir el conjunto de variables a aquellas que realmente tienen un impacto relevante para la modelo de predicción del rendimiento académico, además, esta selección permitirá optimizar recursos computacionales y ahorrar tiempo en la ejecución de los modelos y algoritmos.

Sin embargo, antes de continuar con los procesos de selección, se desplegarán en tablas los detalles de cada variable por cada tipo de reporte, esto, con la finalidad de recapitular toda la información del extenso proceso de análisis que se llevó a cabo, y, adicionalmente,

proporcionarle a usted, respetado lector, identificar aquellas variables que parecen tener una mayor influencia o conexión con el objetivo del análisis.

Para empezar, a continuación, se podrá visualizar las variables identificadas del reporte de asignaturas perdidas:

Tabla 21
Selección de variables del reporte de Asignaturas Perdidas

Variables Relevantes	Variables Irrelevantes	Variables funcionales para futura unión de datos	Variables que ya cumplieron su función	Variable Objetivo
DESC_FACULTAD	VECES	IDD	NOTA_FINAL	APROBADO
DESC_PROGRAMA	PLAN_E	PERIODO		
NIVEL_ESTUDIANTE				
MATE				
CRED				

Por consiguiente, es el turno de presentar las variables identificadas del reporte de calificaciones por carrera:

Tabla 22
Selección de variables del reporte de Calificaciones por Carrera

Variables Relevantes	Variables Irrelevantes	Variables funcionales para futura unión de datos	Variables que ya cumplieron su función
facultad	nota_parcial	id_banner_estudiante	nota_final
id_banner_docente	nota_final_SHATCKG	codigo_periodo	
descripcion_materia			

Finalmente, es momento de mostrar las variables identificadas del reporte de becas:

Tabla 23

Selección de variables del reporte de Becas

Variables Relevantes	Variables Irrelevantes	Variables funcionales para futura unión de datos	Variables que ya cumplieron su función
ETNIA	NACIONALIDAD	ID_BANNER	NOTA_FINAL
NIVEL_POR_MATERIAS	GENERO		
NUM_MATERIAS_SEGUNDA	TIPO_COLEGIO		
_MATRICULA			
NUM_MATERIAS_TERCERA	NIVEL_POR_ASIG_		
_MATRICULA	APROBADAS		
	CATEGORIA_BECA		
	_SOCIOECONOMICA		
	TIENE_SALDO		
	TIENE_PAGO		
	AREA_MATERIAS		
	PROGAMA_DESC		

A partir de la presentación de todas las variables de todos los reportes, es momento de llevar a cabo un proceso de eliminación de las variables irrelevantes y aquellas que han cumplido con su función en el desarrollo de este trabajo de titulación, sin embargo, es necesario mencionar

que la variable APROBADO del reporte de asignaturas perdidas ha sido considerada como la variable objetivo dada su naturaleza y que certifica cuando un estudiante ha reprobado y cuando uno ha aprobado.

Con respecto al proceso de eliminación, en primer lugar, se realizarán los procedimientos sobre las variables del reporte de asignaturas perdidas, el código y la verificación de la eliminación se exhibirá a continuación:

Ilustración 439

Código de la eliminación y verificación de esta de las variables del reporte de Asignaturas Perdidas

```
# Recopilación de las variables por ser eliminadas
por_eliminar = ["VECES", "PLAN_E", "NOTA_FINAL"]
for perdida in perdidas:
    perdida.drop(por_eliminar, axis=1, inplace=True)
```

```
# Verificación de la eliminación
count = 1
for perdida in perdidas:
    print(f"\nAnalizando el reporte perdidas0{count}")
    for column in perdida.columns:
        print(f"\t- {column}")
    count += 1
```

Ilustración 440

Resultados del primer y segundo periodo de la eliminación de variables del reporte de Asignaturas Perdidas

Analizando el reporte perdidas01

- IDD
- PERIODO
- DESC_FACULTAD
- DESC_PROGRAMA
- NIVEL_ESTUDIANTE
- MATE
- APROBADO
- CRED

Analizando el reporte perdidas02

- IDD
- PERIODO
- DESC_FACULTAD
- DESC_PROGRAMA
- NIVEL_ESTUDIANTE
- MATE
- APROBADO
- CRED

Ilustración 441

Resultados del tercer y cuarto periodo de la eliminación de variables del reporte de Asignaturas Perdidas

Analizando el reporte perdidas03

- IDD
- PERIODO
- DESC_FACULTAD
- DESC_PROGRAMA
- NIVEL_ESTUDIANTE
- MATE
- APROBADO
- CRED

Analizando el reporte perdidas04

- IDD
- PERIODO
- DESC_FACULTAD
- DESC_PROGRAMA
- NIVEL_ESTUDIANTE
- MATE
- APROBADO
- CRED

Por otra parte, a continuación, se desplegará el código y verificación de la eliminación de las variables irrelevantes y aquellas que han cumplido su función del reporte de calificaciones por carrera:

Ilustración 442

Código de la eliminación y verificación de esta de las variables del reporte de Calificaciones por Carrera

```
# Recopilación de las variables por ser eliminadas
por_eliminar = ["nota_parcial", "nota_final_SHATCKG", "nota_final"]
for calificacion in calificaciones:
    calificacion.drop(por_eliminar, axis=1, inplace=True)

# Verificación de la eliminación
count = 1
for calificacion in calificaciones:
    print(f"\nAnalizando el reporte calificacion0{count}")
    for column in calificacion.columns:
        print(f"\t- {column}")
    count += 1
```

Ilustración 443

Resultados del primer y segundo periodo de la eliminación de variables del reporte de Calificaciones por Carrera

```
Analizando el reporte calificacion01
- id_banner_estudiante
- codigo_perodo
- facultad
- id_banner_docente
- descripcion_materia
- descripcion_programa

Analizando el reporte calificacion02
- id_banner_estudiante
- codigo_perodo
- facultad
- id_banner_docente
- descripcion_materia
- descripcion_programa
```

Ilustración 444

Resultados del tercer y cuarto periodo de la eliminación de variables del reporte de Calificaciones por Carrera

```
Analizando el reporte calificacion03
- id_banner_estudiante
- codigo_perodo
- facultad
- id_banner_docente
- descripcion_materia
- descripcion_programa

Analizando el reporte calificacion04
- id_banner_estudiante
- codigo_perodo
- facultad
- id_banner_docente
- descripcion_materia
- descripcion_programa
```

Finalmente, a continuación, se podrá visualizar el código y verificación de la eliminación de las variables irrelevantes y de aquellas que su función ha llegado a su fin del reporte de becas:

Ilustración 445

Código de la eliminación y verificación de esta de las variables del reporte de Becas

```
# Recopilación de las variables por ser eliminadas
por_eliminar = [
    "NACIONALIDAD",
    "GENERO",
    "TIPO_COLEGIO",
    "NIVEL_POR_ASIG_APROBADAS",
    "CATEGORIA_BECA_SOCIOECONOMICA",
    "TIENE_PAGO",
    "TIENE_SALDO",
    "PROGAMA_DESC",
    "AREA_MATERIAS",
    "NOTA_FINAL"
]
for beca in becas:
    beca.drop(por_eliminar, axis=1, inplace=True)
```

```
# Verificación de la eliminación
count = 1
for beca in becas:
    print(f"\nAnalizando el reporte becas0{count}")
    for column in beca.columns:
        print(f"\t- {column}")
```

Ilustración 446

Resultados del primer y segundo periodo de la eliminación de variables del reporte de Becas

```
Analizando el reporte becas01
- ID_BANNER
- ETNIA
- NIVEL_POR_MATERIAS
- NUM_MATERIAS_SEGUNDA_MATRICULA
- NUM_MATERIAS_TERCERA_MATRICULA

Analizando el reporte becas01
- ID_BANNER
- ETNIA
- NIVEL_POR_MATERIAS
- NUM_MATERIAS_SEGUNDA_MATRICULA
- NUM_MATERIAS_TERCERA_MATRICULA
```

Ilustración 447

Resultados del tercer y cuarto periodo de la eliminación de variables del reporte de Becas

```
Analizando el reporte becas01
- ID_BANNER
- ETNIA
- NIVEL_POR_MATERIAS
- NUM_MATERIAS_SEGUNDA_MATRICULA
- NUM_MATERIAS_TERCERA_MATRICULA

Analizando el reporte becas01
- ID_BANNER
- ETNIA
- NIVEL_POR_MATERIAS
- NUM_MATERIAS_SEGUNDA_MATRICULA
- NUM_MATERIAS_TERCERA_MATRICULA
```

4.1.14. Vista general de los conjuntos de datos finales

En este corto subcapítulo se podrá admirar los marcos de datos correspondientes a cada periodo de cada reporte resultante, es decir, a continuación, se realizará una demostración enfocada en permitirle a usted, distintivo lector, una vista de los datos resultantes de la fase de comprensión de los datos.

Ilustración 448

Vista del primer periodo del reporte de Asignaturas Perdidas

Periodo 2021-01

```
perdidas[0].head()
```

	IDD	PERIODO	DESC_FACULTAD	DESC_PROGRAMA	NIVEL_ESTUDIANTE	MATE	APROBADO	CRED
0	E5214	202161	Facultad Psicología	Psicología Clínica	09	9	N	6.0
1	E4739	202161	Facultad Cs Humanas	Sociología Política	04	3	N	3.0
2	E10373	202161	Facultad Cs Adm. y Contables	Negocios Internacionales	04	4	N	3.0
3	E2887	202161	Facultad Economía	Economía	04	3	N	4.0
4	E2887	202161	Facultad Economía	Economía	04	3	N	4.0

Ilustración 449

Vista del segundo periodo del reporte de Asignaturas Perdidas

Periodo 2021-02

```
perdidas[1].head()
```

	IDD	PERIODO	DESC_FACULTAD	DESC_PROGRAMA	NIVEL_ESTUDIANTE	MATE	APROBADO	CRED
0	E7055	202166	Facultad Cs Exactas Naturales	Ciencias Biológicas	07	5	N	0.0
1	E10373	202166	Facultad Cs Adm. y Contables	Negocios Internacionales	04	3	N	3.0
2	E2887	202166	Facultad Economía	Economía	06	3	N	4.0
3	E2887	202166	Facultad Economía	Economía	06	3	N	4.0
4	E2887	202166	Facultad Economía	Economía	06	4	N	5.0

Ilustración 450

Vista del tercer periodo del reporte de Asignaturas Perdidas

Periodo 2022-01

```
perdidas[2].head()
```

	IDD	PERIODO	DESC_FACULTAD	DESC_PROGRAMA	NIVEL_ESTUDIANTE	MATE	APROBADO	CRED
0	E3801	202261	Facultad Cs Exactas Naturales	Biología	02	5	N	3.0
1	E3801	202261	Facultad Cs Exactas Naturales	Biología	02	5	N	0.0
2	E3801	202261	Facultad Cs Exactas Naturales	Biología	02	5	N	0.0
3	E7055	202261	Facultad Cs Exactas Naturales	Ciencias Biológicas	08	5	N	2.0
4	E7055	202261	Facultad Cs Exactas Naturales	Ciencias Biológicas	08	5	N	3.0

Ilustración 451

Vista del cuarto periodo del reporte de Asignaturas Perdidas

Periodo 2022-02

```
perdidas[3].head()
```

	IDD	PERIODO	DESC_FACULTAD	DESC_PROGRAMA	NIVEL_ESTUDIANTE	MATE	APROBADO	CRED
0	E3801	202266	Facultad Cs Exactas Naturales	Biología	03	5	N	4.5
1	E4411	202266	Facultad Cs Adm. y Contables	Negocios Internacionales	07	3	N	3.0
2	E4411	202266	Facultad Cs Adm. y Contables	Negocios Internacionales	07	4	N	3.0
3	E10336	202266	Facultad Jurisprudencia	Derecho	08	3	N	4.0
4	E2838	202266	Facultad Cs Adm. y Contables	Negocios Internacionales	07	4	N	0.0

Ilustración 452

Vista del primer periodo del reporte de Calificaciones por Carrera

Periodo 2021-01

```
calificaciones[0].head()
```

	id_banner_estudiante	codigo_peri	facultad	id_banner_docente	descripcion_materia	descripcion_programa
0	E8695	202161	Facultad Cs Adm. y Contables	D936	Cien.Natur.Matemát.y Estadíst.	Administración de Empresas
1	E8695	202161	Facultad Cs Adm. y Contables	D936	Cien.Natur.Matemát.y Estadíst.	Administración de Empresas
2	E8695	202161	Facultad Cs Adm. y Contables	D936	Cien.Natur.Matemát.y Estadíst.	Administración de Empresas
3	E8695	202161	Facultad Cs Adm. y Contables	D936	Cien.Natur.Matemát.y Estadíst.	Administración de Empresas
4	E8695	202161	Facultad Cs Adm. y Contables	D619	Administración	Administración de Empresas

Ilustración 453

Vista del segundo periodo del reporte de Calificaciones por Carrera

Periodo 2021-02

```
calificaciones[1].head()
```

	id_banner_estudiante	codigo_peri	facultad	id_banner_docente	descripcion_materia	descripcion_programa
0	E8695	202166	Facultad Cs Adm. y Contables	D874	Administración	Administración de Empresas
1	E8695	202166	Facultad Cs Adm. y Contables	D874	Administración	Administración de Empresas
2	E8695	202166	Facultad Cs Adm. y Contables	D874	Administración	Administración de Empresas
3	E8695	202166	Facultad Cs Adm. y Contables	D874	Administración	Administración de Empresas
4	E8695	202166	Facultad Cs Adm. y Contables	D619	Administración	Administración de Empresas

Ilustración 454

Vista del tercer periodo del reporte de Calificaciones por Carrera

Periodo 2022-01

```
calificaciones[2].head()
```

	id_banner_estudiante	codigo_perodo	facultad	id_banner_docente	descripcion_materia	descripcion_programa
0	E8695	202261	Facultad Cs Adm. y Contables	D161	Administración	Administración de Empresas
1	E8695	202261	Facultad Cs Adm. y Contables	D161	Administración	Administración de Empresas
2	E8695	202261	Facultad Cs Adm. y Contables	D161	Administración	Administración de Empresas
3	E8695	202261	Facultad Cs Adm. y Contables	D161	Administración	Administración de Empresas
4	E8695	202261	Facultad Cs Adm. y Contables	D782	Cien.Natur.Matemát.y Estadíst.	Administración de Empresas

Ilustración 455

Vista del cuarto periodo del reporte de Calificaciones por Carrera

Periodo 2022-02

```
calificaciones[3].head()
```

	id_banner_estudiante	codigo_perodo	facultad	id_banner_docente	descripcion_materia	descripcion_programa
0	E3766	202266	Facultad Com Lingüística y Lit	D638	Artes y Humanidades	Fisioterapia
1	E3766	202266	Facultad Com Lingüística y Lit	D638	Artes y Humanidades	Fisioterapia
2	E3766	202266	Facultad Com Lingüística y Lit	D638	Artes y Humanidades	Fisioterapia
3	E3766	202266	Facultad Com Lingüística y Lit	D638	Artes y Humanidades	Fisioterapia
4	E3766	202266	Facultad Cs Exactas Naturales	D672	Investigación	Fisioterapia

Ilustración 456

Vista del primer periodo del reporte de Becas

Periodo 2021-01

```
becas[0].head()
```

	ID_BANNER	ETNIA	NIVEL_POR_MATERIAS	NUM_MATERIAS_SEGUNDA_MATRICULA	NUM_MATERIAS_TERCERA_MATRICULA
0	E9532	NaN	01	1	0
1	E2371	Mestizo	07	0	0
2	E2838	NaN	04	1	0
3	E3781	Mestizo	10	0	0
4	E2249	NaN	02	0	0

Ilustración 457

Vista del segundo periodo del reporte de Becas

Periodo 2021-02

```
becas[1].head()
```

	ID_BANNER	ETNIA	NIVEL_POR_MATERIAS	NUM_MATERIAS_SEGUNDA_MATRICULA	NUM_MATERIAS_TERCERA_MATRICULA
0	E2371	Mestizo	08	0	0
1	E2838	NaN	05	0	0
2	E2249	NaN	03	0	0
3	E11819	Otro	07	0	0
4	E2739	NaN	05	0	0

Ilustración 458

Vista del tercer periodo del reporte de Becas

Periodo 2022-01

```
becas[2].head()
```

	ID_BANNER	ETNIA	NIVEL_POR_MATERIAS	NUM_MATERIAS_SEGUNDA_MATRICULA	NUM_MATERIAS_TERCERA_MATRICULA
0	E2838	NaN	06	0	0
1	E5194	NaN	01	0	0
2	E1659	Mestizo	07	1	0
3	E11819	Otro	08	0	0
4	E2739	NaN	06	0	0

Ilustración 459

Vista del cuarto periodo del reporte de Becas

Periodo 2022-02

```
becas[3].head()
```

	ID_BANNER	ETNIA	NIVEL_POR_MATERIAS	NUM_MATERIAS_SEGUNDA_MATRICULA	NUM_MATERIAS_TERCERA_MATRICULA
0	E2838	NaN	07	0	0
1	E5194	NaN	02	0	0
2	E2739	NaN	07	0	0
3	E7159	NaN	03	0	0
4	E7259	NaN	05	0	0

De manera concluyente, se puede admirar claramente los resultados y las razones del extenso proceso de la fase de entendimiento de los datos.

4.1.15. Exportación de los conjuntos de datos para la siguiente fase

En penúltimo término con respecto a la fase de entendimiento de los datos, es momento de realizar la exportación de los datos procesados en el capítulo, esto, debido a que el proceso de preparación de los datos se realizó en un documento diferente del que se realizó la presente comprensión de los datos, cabe resaltar que el término “pre-preparados” usado en el código es un término que ha sido adaptado para el contexto de la ciencia de datos, el concepto empleado es en síntesis, todo lo realizado hasta el momento antes de preparar los datos, es decir, el resultado de la fase de comprensión de los datos. El código empleado para la exportación es el siguiente:

Ilustración 460

Código de exportación de los datos pre-preparados

```
# Variables de directorio y reportes deseados
directorio = "../Datos Estudiantes - Pre-Preparados/"
nombres_reportes = [
    "Reporte de asignaturas perdidas por periodo y reincidencia ",
    "Reporte de Calificaciones por Carrera ",
    "Reporte de detalle de estudiantes inscritos con beca "
]

# Si no existe el directorio se crea
if not os.path.exists(directorio):
    os.makedirs(directorio)

# Recopilación de todos los reportes
reportes = [perdidas, calificaciones, becas]

# Asignación de nuevos nombres a los reportes
count = 0
periodos = ["2021-01", "2021-02", "2022-01", "2022-02"]
for reporte in reportes:
    for periodo in reporte:
        periodo.name = periodos[count]
        count += 1
    count = 0

# Exportación de los reportes y sus periodos
for i in range(len(reportes)):
    for reporte in reportes[i]:
        reporte.to_csv(directorio + nombres_reportes[i] + reporte.name + ".csv", index=False, encoding="utf-8")

print("Se han creado los archivos con éxito!")

Se han creado los archivos con éxito!
```


4.1.16. Planteamiento de la hipótesis para el modelo de Machine Learning

Finalmente se ha logrado alcanzar el último escalón de la fase de comprensión de los datos, por lo que, tras el extenso análisis y procesos de selección de variables, se llegó al consenso de la siguiente hipótesis:

Las variables referidas a la facultad a la que pertenece un estudiante, su carrera, su nivel académico, el área de asignatura que cursará o está considerando cursar, la cantidad de créditos que equivale la asignatura, el docente que imparte dicha asignatura, la etnia, su cantidad de asignaturas con segunda y tercera matrícula influyen directamente en el rendimiento académico de los estudiantes universitarios en la Pontificia Universidad Católica del Ecuador.

4.2. Preparación de los datos

La fase de preparación de los datos, como se ha mencionado en el capítulo 2, es la fase que corresponde a todas las actividades que serán llevadas a cabo sobre los datos obtenidos con la finalidad de construir un conjunto de datos final con todos los datos con los que será entrenado el modelo de predicción de rendimiento académico. (Chapman, y otros, 1999, pág. 14). Es decir, se aplicarán técnicas como la limpieza de datos, la construcción o formateo de datos, la integración de datos en el caso de presentar múltiples fuentes de datos y la normalización y estandarización. (Hotz, 2023).

4.2.1. Librerías requeridas

En esta ocasión, se utilizará nuevamente pandas y numpy para el manejo de los datos y valores faltantes, adicionalmente también se utilizará la librería “os” para poder ubicar los archivos deseados, finalmente, se emplearán los módulos “MinMaxScaler” con el objetivo de transformar características al escalar cada una a un determinado rango. (scikitlearn, s.f.) y

“StandardScaler” para estandarizar características al eliminar la media y escalar la varianza de esta. (scikitlearn, s.f.). Finalmente, se utilizó “category_encoders” que es “A set of scikit-learn-style transformers for encoding categorical variables into numeric with different techniques.” [Un conjunto de transformadores de estilo scikit-learn para codificar variables categóricas en numéricas con diferentes técnicas.] (Category Encoders, 2022)

Ilustración 461

Librerías requeridas para la fase de preparación de los datos

Librerías requeridas

```
# Librerías para el manejo de los datos
import pandas as pd
import numpy as np

# Permite importar los conjuntos de datos
import os

# Librerías para la normalización y estandarización
from sklearn.preprocessing import MinMaxScaler, StandardScaler
import category_encoders as ce
```

4.2.2. Carga de los conjuntos de datos

Como primer paso para la preparación de los datos, se debe importar los mismos, sin embargo, para esto se realizó un proceso similar al realizado en la fase de comprensión de los datos, por lo que, igualmente es necesario comprender que estos se importarán como marcos de datos bajo la nomenclatura visible en la siguiente imagen:

Ilustración 462

Nomenclatura de los marcos de datos de los conjuntos de datos obtenidos

Reporte de asignaturas perdidas por periodo y reincidencia:

- Reporte de asignaturas perdidas por periodo y reincidencia 2021-01 -> perdidas01
- Reporte de asignaturas perdidas por periodo y reincidencia 2021-02 -> perdidas02
- Reporte de asignaturas perdidas por periodo y reincidencia 2022-01 -> perdidas03
- Reporte de asignaturas perdidas por periodo y reincidencia 2022-02 -> perdidas04

Reporte de Calificaciones por Carrera:

- Reporte de Calificaciones por Carrera 2021-01 -> calificaciones01
- Reporte de Calificaciones por Carrera 2021-02 -> calificaciones02
- Reporte de Calificaciones por Carrera 2022-01 -> calificaciones03
- Reporte de Calificaciones por Carrera 2022-02 -> calificaciones04

Reporte de detalle de estudiantes inscritos con beca:

- Reporte de detalle de estudiantes inscritos con beca 2021-01 -> beca01
- Reporte de detalle de estudiantes inscritos con beca 2021-02 -> beca02
- Reporte de detalle de estudiantes inscritos con beca 2022-01 -> beca03
- Reporte de detalle de estudiantes inscritos con beca 2022-02 -> beca04

A partir de la nomenclatura visible, es importante mencionar que los diferentes conjuntos de datos que se importarán, además de corresponder a la nomenclatura descrita, estos se recopilaran en listas individuales por reporte, es decir, todos los conjuntos de datos correspondientes a cada reporte serán almacenados en listas específicas de cada reporte. Para los periodos del reporte de asignaturas perdidas se considerará una lista llamada “perdidas”, para los periodos del reporte de calificaciones por carrera se considerará una lista llamada “calificaciones”, y, finalmente, para los periodos del reporte de becas se considerará una lista llamada “becas”.

El código utilizado para la importación de los conjuntos de datos y el resultado de este son los siguiente:

Ilustración 463

Función para la importación de conjuntos de datos y colocación de nombres dada una nomenclatura específica

```
def get_reportes(directory, reporte_deseado):  
    """  
    La función evaluará un directorio dado y devolverá dataframes en función de los reportes deseados.  
    Si la función no encuentra ningún reporte deseado devolverá una lista vacía.  
    """  
  
    # Obtenemos la lista de archivos en el directorio  
    files = os.listdir(directory)  
  
    # Creamos una variable que almacenará los reportes como dataframes  
    dataframes = []  
  
    # Se itera en busca de los archivos requeridos  
    for file in files:  
        # Variable con la dirección del archivo evaluado  
        filepath = os.path.join(directory, file)  
        # Evaluamos si es el reporte deseado  
        if os.path.isfile(filepath) and " ".join(file.split()[:-1]) == reporte_deseado:  
            # Se extrae el archivo CSV como dataframe  
            temp_df = pd.read_csv(filepath, sep=";", low_memory=False)  
            # Se agrega el dataframe a la lista con todos los dataframes.  
            dataframes.append(temp_df)  
  
    # Se retorna la lista con todos los reportes como dataframes  
    return dataframes
```

Ilustración 464

Código para la importación de los datos en marcos de datos para su uso en la fase de preparación de datos

```
# Variables de directorio y reportes deseados
directorio = "../Datos Estudiantes - Pre-Preparados/"
reportes_deseados = [
    "Reporte de asignaturas perdidas por periodo y reincidencia",
    "Reporte de Calificaciones por Carrera",
    "Reporte de detalle de estudiantes inscritos con beca"
]

# Extraemos los reportes de asignaturas perdidas
perdidas = get_reportes(directorio, reportes_deseados[0])
# Extraemos los reportes de calificaciones
calificaciones = get_reportes(directorio, reportes_deseados[1])
# Extraemos los reportes de becas
becas = get_reportes(directorio, reportes_deseados[2])

# Damos nombres a los dataframes
for i in range(4):
    perdidas[i].name = "perdidas0" + str(i+1)
    calificaciones[i].name = "calificaciones0" + str(i+1)
    becas[i].name = "becas0" + str(i+1)

# Verificación de nombres
print("Los reportes importados son:")
print("\nReportes de asignaturas perdidas: ")
for perdida in perdidas: print(f"\t- {perdida.name}")
print("\nReportes de calificaciones por carrera: ")
for calificacion in calificaciones: print(f"\t- {calificacion.name}")
print("\nReportes de becas: ")
for beca in becas: print(f"\t- {beca.name}")
```

Ilustración 465

Resultados de la importación de los datos

```
Los reportes importados son:

Reportes de asignaturas perdidas:
- perdidas01
- perdidas02
- perdidas03
- perdidas04

Reportes de calificaciones por carrera:
- calificaciones01
- calificaciones02
- calificaciones03
- calificaciones04

Reportes de becas:
- becas01
- becas02
- becas03
- becas04
```

4.2.3. Revisión inicial de los conjuntos de datos

Una vez importado los datos, es necesario realizar una revisión inicial de los mismos, es decir, es necesario recordar el sentido, significado y esencia de los datos que se poseen. Para lograr este acometido, a continuación, se podrán encontrar tablas con la descripción de cada variable que se encuentra en el conjunto de datos pre-preparado de la fase comprensión de los datos. Para empezar, se desplegarán las variables del reporte de asignaturas perdidas:

Tabla 24

Revisión del Reporte de Asignaturas Perdidas

Característica	Descripción
IDD	Identificador único del estudiante en el banner.
PERIODO	Periodo académico en el que se perdió la asignatura.
DESC_FACULTAD	Nombre completo de la facultad a la que pertenece el estudiante.
DESC_PROGRAMA	Nombre del programa en el que se encuentra inscrito el estudiante.

NIVEL_ESTUDIANTE	Nivel académico del estudiante dentro del programa.
MATE	Código del área a la que pertenece la asignatura.
APROBADO	Indicador si el estudiante aprobó o reprobó el curso.
CRED	Número de créditos asignados al curso.

Dentro de las variables que disponen, es imperativo notar que se encuentra en código, esto implica la necesidad de presentar un diccionario de cada código con su significado:

Tabla 25
Diccionario de códigos de áreas de asignaturas

Có- digo	Descripción
1	Educación
2	Artes y Humanidades
3	Cien.Soc.period.form y Derecho
4	Administración
5	Cien.Natur.Matemát.y Estadíst.
6	Tecno.Informac.y Comun.(TIC)
7	Ingeniería Industria y Constr.
8	Agricul.Silvic.pesca y veteri.
9	Salud y Bienestar
10	Servicios
11	Investigación

Por otra parte, es momento de presentar las variables del reporte de calificaciones por carrera:

Tabla 26

Revisión del Reporte de Calificaciones por Carrera

Característica	Descripción
id_banner_estudiante	Identificador único del estudiante en el sistema Banner.
codigo_periodo	Código del período académico en el que se registraron las calificaciones.
facultad	Nombre de la facultad a la que pertenece el estudiante.
id_banner_docente	Identificador único del docente en el sistema de gestión académica.
descripcion_materia	Área de la materia cursada.
descripcion_programa	Nombre del programa en el que se encuentra inscrito el estudiante.

Finalmente, ha llegado la hora de presentar las variables del reporte de becas:

Tabla 27

Revisión del Reporte de Becas

Característica	Descripción
ID_BANNER	Identificador único del estudiante en el sistema Banner.
ETNIA	Etnia a la que pertenece el estudiante.
NIVEL_POR_MATERIAS	Nivel académico del estudiante.
NUM_MATERIAS_SEGUNDA_MATRICULA	Número de materias en la segunda matrícula del estudiante.
NUM_MATERIAS_TERCERA_MATRICULA	Número de materias en la tercera matrícula del estudiante.

Una vez se ha permitido la visualización de la descripción de todas las variables que se poseen para esta fase de preparación de los datos, se puede realizar una vista general de los

misimos, sin embargo, dado que ya se realizó esta vista en el anterior capítulo, en el presenta, únicamente se desplegarán los primeros periodos de cada reporte:

Ilustración 466

Vista del primer periodo del reporte de Asignaturas Perdidas

Periodo 2021-01

```
perdidas[0].head()
```

	IDD	PERIODO	DESC_FACULTAD	DESC_PROGRAMA	NIVEL_ESTUDIANTE	MATE	APROBADO	CRED
0	E5214	202161	Facultad Psicología	Psicología Clínica	09	9	N	6.0
1	E4739	202161	Facultad Cs Humanas	Sociología Política	04	3	N	3.0
2	E10373	202161	Facultad Cs Adm. y Contables	Negocios Internacionales	04	4	N	3.0
3	E2887	202161	Facultad Economía	Economía	04	3	N	4.0
4	E2887	202161	Facultad Economía	Economía	04	3	N	4.0

Ilustración 467

Vista del primer periodo del reporte de Calificaciones por Carrera

Periodo 2021-01

```
calificaciones[0].head()
```

	id_banner_estudiante	codigo_peri	facultad	id_banner_docente	descripcion_materia	descripcion_programa
0	E8695	202161	Facultad Cs Adm. y Contables	D936	Cien.Natur.Matemát.y Estadíst.	Administración de Empresas
1	E8695	202161	Facultad Cs Adm. y Contables	D936	Cien.Natur.Matemát.y Estadíst.	Administración de Empresas
2	E8695	202161	Facultad Cs Adm. y Contables	D936	Cien.Natur.Matemát.y Estadíst.	Administración de Empresas
3	E8695	202161	Facultad Cs Adm. y Contables	D936	Cien.Natur.Matemát.y Estadíst.	Administración de Empresas
4	E8695	202161	Facultad Cs Adm. y Contables	D619	Administración	Administración de Empresas

Ilustración 468

Vista del primer periodo del reporte de Becas

Periodo 2021-01

```
becas[0].head()
```

	ID_BANNER	ETNIA	NIVEL_POR_MATERIAS	NUM_MATERIAS_SEGUNDA_MATRICULA	NUM_MATERIAS_TERCERA_MATRICULA
0	E9532	NaN	01	1	0
1	E2371	Mestizo	07	0	0
2	E2838	NaN	04	1	0
3	E3781	Mestizo	10	0	0
4	E2249	NaN	02	0	0

4.2.4. Limpieza inicial de datos

Dentro de todo proceso de ciencia de datos, y en el contexto del desarrollo de un modelo de predicción del rendimiento académico, la limpieza de datos resulta un proceso crítico ya que es posible la existencia de una variedad de errores en los datos que se están tratando que pueden resultar afectando al modelo de predicción deseado. (Amazon, s.f.). Sin embargo, cuando se refiere el nombre a “limpieza inicial” se refiere a que esta etapa no se dedicará a ejercer todos los pasos que comúnmente se llevan a cabo en una limpieza apropiada, para ello se encuentra el subcapítulo 4.2.7, no obstante, en esta sección si se realizará un procedimiento de la limpieza, esta es la eliminación de valores duplicados.

Se ha tomado la decisión de realizar un inicial proceso de eliminación de duplicados debido a que los datos duplicados tiene la posibilidad de introducir una serie de problemas tanto para el modelo de machine learning como para el proceso de preparación de los datos debido a que consume espacio de almacenamiento y desordena en gran medida los datos. (Ibarrera, 2020) Por lo que, se ha tomado la decisión de realizar este proceso con la finalidad de mejorar la fluidez del proceso de unificación de los de datos obtenidos.

Para empezar, se realizará el proceso de eliminación de duplicados en el reporte de asignaturas perdidas, este se llevó a cabo mediante el siguiente código:

Ilustración 469

Código para la eliminación de duplicados del reporte de Asignaturas Perdidas

```
# Visualización de Las dimensiones antes de la eliminación de duplicados
for perdida in perdidas:
    print(f"Existen {perdida.shape[0]} registros y {perdida.shape[1]} variables.")
```

```
Existen 1670 registros y 8 variables.
Existen 2104 registros y 8 variables.
Existen 3069 registros y 8 variables.
Existen 3141 registros y 8 variables.
```

```
# Se procede a iterar sobre todos los periodos
for perdida in perdidas:
    # Se eliminan los duplicados del periodo
    perdida.drop_duplicates(inplace=True)
```

```
# Visualización de Las dimensiones después de la eliminación de duplicados
for perdida in perdidas:
    print(f"Existen {perdida.shape[0]} registros y {perdida.shape[1]} variables.")
```

```
Existen 1292 registros y 8 variables.
Existen 1668 registros y 8 variables.
Existen 2518 registros y 8 variables.
Existen 2640 registros y 8 variables.
```

Por otra parte, se continuo con el proceso de eliminación de duplicados con el reporte de calificaciones por carrera con el siguiente código:

Ilustración 470

Código para la eliminación de duplicados del reporte de Calificaciones por Carrera

```
# Visualización de Las dimensiones antes de la eliminación de duplicados
for calificacion in calificaciones:
    print(f"Existen {calificacion.shape[0]} registros y {calificacion.shape[1]} variables.")
```

```
Existen 163810 registros y 6 variables.
Existen 158818 registros y 6 variables.
Existen 150108 registros y 6 variables.
Existen 142205 registros y 6 variables.
```

```
# Se procede a iterar sobre todos los periodos
for calificacion in calificaciones:
    # Se eliminan los duplicados del periodo
    calificacion.drop_duplicates(inplace=True)
```

```
# Visualización de Las dimensiones después de la eliminación de duplicados
for calificacion in calificaciones:
    print(f"Existen {calificacion.shape[0]} registros y {calificacion.shape[1]} variables.")
```

```
Existen 39861 registros y 6 variables.
Existen 39001 registros y 6 variables.
Existen 36703 registros y 6 variables.
Existen 35056 registros y 6 variables.
```

Finalmente, se concluyó el proceso con la eliminación de los duplicados del reporte de Becas, el código puede ser visualizado a continuación:

Ilustración 471

Código para la eliminación de duplicados del reporte de Calificaciones por Carrera

```
# Visualización de las dimensiones antes de la eliminación de duplicados
for beca in becas:
    print(f"Existen {beca.shape[0]} registros y {beca.shape[1]} variables.")
```

```
Existen 9399 registros y 5 variables.
Existen 8564 registros y 5 variables.
Existen 8336 registros y 5 variables.
Existen 8255 registros y 5 variables.
```

```
# Se procede a iterar sobre todos los periodos
for beca in becas:
    # Se eliminan los duplicados del periodo
    beca.drop_duplicates(inplace=True)
```

```
# Visualización de las dimensiones después de la eliminación de duplicados
for beca in becas:
    print(f"Existen {beca.shape[0]} registros y {beca.shape[1]} variables.")
```

```
Existen 8608 registros y 5 variables.
Existen 8562 registros y 5 variables.
Existen 8240 registros y 5 variables.
Existen 8255 registros y 5 variables.
```

En conclusión, el proceso de eliminación llevado a cabo dejó los siguientes resultados:

Tabla 28

Resultados de la Eliminación de Duplicados del Reporte de Asignaturas Perdidas

Periodo	Cantidad inicial de registros	Cantidad de registros eliminados	Cantidad actual de registros
2021-01	1670	378	1292
2021-02	2104	436	1668
2022-01	3069	551	2518
2022-02	3141	501	2640

Tabla 29

Resultados de la Eliminación de Duplicados del Reporte de Calificaciones por Carrera

Periodo	Cantidad inicial de registros	Cantidad de registros eliminados	Cantidad actual de registros
2021-01	163810	123949	39861
2021-02	158818	119817	39001
2022-01	150108	113405	36703
2022-02	142205	107149	35056

Tabla 30

Resultados de la Eliminación de Duplicados del Reporte de Becas

Periodo	Cantidad inicial de registros	Cantidad de registros eliminados	Cantidad actual de registros
2021-01	9399	791	8608
2021-02	8564	2	8562
2022-01	8336	96	8240
2022-02	8255	0	8255

4.2.5. Integración de los conjuntos de datos

Una vez finalizado el proceso de limpieza inicial, en este subcapítulo se realizará la unión de todos los conjuntos de datos con la finalidad de crear un único conjunto de datos con el que procederá a la fase de modelado, sin embargo, para realizar la unificación es necesario tener en cuenta diversos aspectos.

Metodología de unificación

Los tres reportes de datos poseen variables clave que facilitan su integración en un solo conjunto de datos como el identificador del estudiante (IDD en "Asignaturas Perdidas",

id_banner_estudiante en "Calificaciones por Carrera" y ID_BANNER en "Becas"), esta, es una variable única para cada estudiante, proporcionando un enlace natural entre los diferentes conjuntos de datos.

Sin embargo, a pesar de que estos conjuntos de datos proporcionan información complementaria, se debe tener presente que no todos los estudiantes figuran en todos ellos, es decir, solo los estudiantes que perdieron asignaturas aparecerán en "Asignaturas Perdidas", y únicamente los estudiantes beneficiarios de una beca socioeconómica estarán en "Becas"; de esta manera, durante la unificación de los conjuntos de datos, es probable que surjan datos faltantes en algunas columnas para ciertos estudiantes.

Comportamiento esperado tras la unificación

Una vez unificados los conjuntos de datos, se espera obtener una representación más completa de la experiencia académica de cada estudiante, cada registro en el nuevo conjunto de datos unificado representará a un estudiante único, incluyendo toda la información disponible sobre sus asignaturas, calificaciones, carrera, facultad, nivel académico, etnia, cantidad de asignaturas con segunda y tercera matrícula, y estatus de beca. Esta diversidad de datos permitirá realizar un análisis más exhaustivo y preciso, facilitando la verificación de la hipótesis planteada.

Estrategia de Integración de los Conjuntos de Datos de Reportes Académicos de la Pontificia Universidad Católica del Ecuador

Antes de la unificación de los conjuntos de datos, se debe establecer una estrategia clara y precisa para prevenir la pérdida de información relevante o la introducción de errores, por lo que, a continuación, se describe la metodología propuesta para integrar estos conjuntos de datos:

Preprocesamiento de los datos. Para garantizar que la integración de los datos se realice sin contratiempos, es esencial realizar un preprocesamiento de los datos, el mismo, consistirá en verificar y asegurar la compatibilidad entre las variables en los diferentes conjuntos de datos.

Unificación de los conjuntos de datos por cada periodo. La estrategia clave en este paso es unificar los conjuntos de datos realizando una fusión de tipo "outer" basada en la identificación del estudiante, lo que asegurará que todos los registros únicos en los tres conjuntos de datos se mantengan y no se pierdan. No obstante, hay que tener en cuenta que en aquellos casos donde un registro específico no se encuentre en un conjunto de datos, las columnas correspondientes se llenarán con valores nulos, sin embargo, este aspecto se manejará en la etapa de limpieza de datos avanzada.

El objetivo de esta sección es unir los conjuntos de datos de cada reporte dado el periodo, es decir, los conjuntos de datos del periodo 2021-01, después los del 2021-02, etc.

Preprocesamiento de los conjuntos de datos de los periodos. Antes de unir los conjuntos de datos de los diferentes periodos académicos, es crucial realizar un segundo preprocesamiento, lo que garantizará que los datos de cada periodo unificado sean compatibles entre sí, de manera que puedan integrarse sin introducir inconsistencias o errores.

Unificación de los conjuntos de datos de los periodos. Finalmente, después de haber unificado los conjuntos de datos dado el periodo y haberlos preprocesado, el siguiente paso será combinar todos estos conjuntos de datos en uno solo, es decir, realizar la unificación de todos los periodos. El resultado esperado será un conjunto de datos consolidado que contenga información de todos los periodos académicos, manteniendo la estructura y la coherencia requerida para el modelo de machine learning.

Preprocesamiento de los datos

Como se ha mencionado, para garantizar que la integración de los datos se realice sin contratiempos, es esencial realizar un preprocesamiento de los datos, el mismo, consistirá en verificar y asegurar la compatibilidad entre las variables en los diferentes conjuntos de datos. Para esto, primero se realizará una breve identificación de los nombres de las variables de los reportes debido a que para realizar la unión se deben compartir nombres semejantes, además, permitirá un mejor entendimiento de los datos, el código y sus resultados podrán ser identificados a continuación:

Ilustración 472

Código para la identificación de las variables de los reportes

```
# Se toma todas las variables presentes en los tres reportes
variables_perdidas = [variable for variable in perdidas[0].columns]
variables_calificaciones = [variable for variable in calificaciones[0].columns]
variables_becas = [variable for variable in becas[0].columns]

# Impresión de las variables
print("Variables en el reporte de asignaturas perdidas:")
for variable in variables_perdidas:
    print(f"\t-{variable}")
print("\nVariables en el reporte de calificaciones por carrera:")
for variable in variables_calificaciones:
    print(f"\t-{variable}")
print("\nVariables en el reporte de becas:")
for variable in variables_becas:
    print(f"\t-{variable}")
```


Ilustración 473

Identificación de las variables de los reportes

Variables en el reporte de asignaturas perdidas:

- IDD
- PERIODO
- DESC_FACULTAD
- DESC_PROGRAMA
- NIVEL_ESTUDIANTE
- MATE
- APROBADO
- CRED

Variables en el reporte de calificaciones por carrera:

- id_banner_estudiante
- codigo_periodes
- facultad
- id_banner_docente
- descripcion_materia
- descripcion_programa

Variables en el reporte de becas:

- ID_BANNER
- ETNIA
- NIVEL_POR_MATERIAS
- NUM_MATERIAS_SEGUNDA_MATRICULA
- NUM_MATERIAS_TERCERA_MATRICULA

Como se puede observar, existen varias variables que se refiere al mismo tipo de datos, como, por ejemplo, las variables IDD, id_banner_estudiante e ID_BANNER. Por lo que, se evidencia la necesidad de realizar una armonización de los nombres de estas variables, no obstante, la armonización también se llevará a cabo en aquellas variables que no poseen un semejante en los demás reportes, esto con la finalidad de asegurar que todas las variables posean un formato legible y coherente. Las consideraciones para la armonización de variables se detallarán en las siguientes tablas:

Tabla 31

Consideraciones para la armonización de las variables del reporte de Asignaturas Perdidas

Antiguo nombre	Nombre armonizado
IDD	estudiante

PERIODO	periodo
DESC_FACULTAD	facultad
DESC_PROGRAMA	carrera
NIVEL	nivel
MATE	area_asignatura
APROBADO	aprobado
CRED	creditos

Tabla 32

Consideraciones para la armonización de las variables del reporte de Calificaciones por Carrera

Antiguo nombre	Nombre armonizado
id_banner_estudiante	estudiante
codigo_periodo	periodo
facultad	facultad
id_banner_docente	docente
descripcion_materia	area_asignatura
descripcion_programa	carrera

Tabla 33

Consideraciones para la armonización de las variables del reporte de Becas

Antiguo nombre	Nombre armonizado
ID_BANNER	estudiante
ETNIA	etnia
NIVEL_POR_MATERIAS	nivel
NUM_MATERIAS_SEGUNDA_MATRICULA	segundas_matriculas
NUM_MATERIAS_TERCERA_MATRICULA	terceras_matriculas

A partir de las consideraciones identificadas en las tablas anteriores, se llevará a cabo el proceso de armonización con la finalidad de que los nombres de las variables de cada reporte correspondan a las consideraciones mencionadas. El código para la realización de la armonización se puede presenciar a continuación:

Ilustración 474

Definición de diccionarios para el cambio de nombre de las variables de los reportes

```
# Definición de un diccionarios con Los nombres de las variables antiguos y armonizados
nuevas_perdidas = {
    "IDD" : "estudiante",
    "PERIODO" : "periodo",
    "DESC_FACULTAD" : "facultad",
    "DESC_PROGRAMA" : "carrera",
    "NIVEL_ESTUDIANTE" : "nivel",
    "MATE" : "area_asignatura",
    "APROBADO" : "aprobado",
    "CRED" : "creditos"
}

# Definición de un diccionarios con Los nombres de las variables antiguos y armonizados
nuevas_calificaciones = {
    "id_banner_estudiante" : "estudiante",
    "codigo_periodo" : "periodo",
    "facultad" : "facultad",
    "id_banner_docente" : "docente",
    "descripcion_materia" : "area_asignatura",
    "descripcion_programa" : "carrera"
}

# Definición de un diccionarios con Los nombres de las variables antiguos y armonizados
nuevas_becas = {
    "ID_BANNER" : "estudiante",
    "ETNIA" : "etnia",
    "NIVEL_POR_MATERIAS" : "nivel",
    "NUM_MATERIAS_SEGUNDA_MATRICULA" : "segundas_matriculas",
    "NUM_MATERIAS_TERCERA_MATRICULA" : "terceras_matriculas"
}
```

Finalmente, una vez ejecuta la armonización, se procede a identificar los nuevos nombres que poseen las variables de los distintos conjuntos de datos:

Ilustración 475

Código de la ejecución de la armonización de los nombres de las variables

```
# Cambio de nombres de Las variables de Los reportes
for perdida in perdidas:
    perdida.rename(columns=nuevas_perdidas, inplace=True)

# Cambio de nombres de Las variables de Los reportes
for calificacion in calificaciones:
    calificacion.rename(columns=nuevas_calificaciones, inplace=True)

# Cambio de nombres de Las variables de Los reportes
for beca in becas:
    beca.rename(columns=nuevas_becas, inplace=True)
```

Ilustración 476

Identificación de variables de los reportes tras la armonización

```
# Se toma todas las variables presentes en los tres reportes
variables_perdidas = [variable for variable in perdidas[0].columns]
variables_calificaciones = [variable for variable in calificaciones[0].columns]
variables_becas = [variable for variable in becas[0].columns]

# Impresión de las variables
print("Variables en el reporte de asignaturas perdidas:")
for variable in variables_perdidas:
    print(f"\t-{variable}")
print("\nVariables en el reporte de calificaciones por carrera:")
for variable in variables_calificaciones:
    print(f"\t-{variable}")
print("\nVariables en el reporte de becas:")
for variable in variables_becas:
    print(f"\t-{variable}")
```

Ilustración 477

Identificación de los nombres de las variables armonizadas

Variables en el reporte de asignaturas perdidas:

- estudiante
- periodo
- facultad
- carrera
- nivel
- area_asignatura
- aprobado
- creditos

Variables en el reporte de calificaciones por carrera:

- estudiante
- periodo
- facultad
- docente
- area_asignatura
- carrera

Variables en el reporte de becas:

- estudiante
- etnia
- nivel
- segundas_matriculas
- terceras_matriculas

En conclusión, se ha logrado armonizar los nombres de las variables de los distintos reportes, el siguiente paso corresponde a identificar a las variables que se atribuyen a los mismos datos, este proceso iniciará con el despliegue de las variables del reporte de asignaturas perdidas en relación con el reporte calificaciones por carrera y el reporte becas:

Tabla 34

Identificación de variables que se atribuyen a los mismos datos del reporte de Asignaturas Perdidas

Calificaciones por Carrera	Becas
estudiante	estu- diente
periodo	nivel
facultad	

carrera

area_asignatura

En segundo término, se podrán visualizar las variables del reporte de calificaciones por carrera que se atribuyen a los mismos datos del reporte de asignaturas perdidas y del reporte de becas:

Tabla 35

Identificación de variables que se atribuyen a los mismos datos del reporte de Calificaciones por Carrera

Asignaturas perdidas	Becas
estudiante	estudiante
periodo	
facultad	
carrera	
area_asignatura	

Finalmente, se mostrarán las variables del reporte de becas que se atribuyen a los mismos datos del reporte de asignaturas perdidas y del reporte de calificaciones por carrera:

Tabla 36

Identificación de variables que se atribuyen a los mismos datos del reporte de Becas

Asignaturas perdi- das	Calificaciones por Carrera
estudiante	estudiante
nivel	

Teniendo este presente, se realizará un análisis de compatibilidad las variables identificadas que se atribuyen a los mismos datos, este análisis de compatibilidad se lo realizará en términos del tipo de dato de los datos. El análisis comenzará con la evaluación de la compatibilidad del primer periodo, es decir, se analizará el primer periodo de los tres reportes datos, para ello, se realizará una simple inspección visual:

Ilustración 478

Comparativa de tipos de datos del primer periodo de los datos

Periodo 2021-01

perdidas[0].dtypes

estudiante	object
periodo	int64
facultad	object
carrera	object
nivel	object
area_asignatura	int64
aprobado	object
creditos	float64
dtype:	object

calificaciones[0].dtypes

estudiante	object
periodo	int64
facultad	object
docente	object
area_asignatura	object
carrera	object
dtype:	object

becas[0].dtypes

estudiante	object
etnia	object
nivel	object
segundas_matriculas	int64
terceras_matriculas	int64
dtype:	object

A partir de la comparativa se puede concluir que la variable `area_asignatura` en el reporte de asignaturas perdidas es de tipo diferente del reporte de calificaciones por carrera, por lo que, las variables que se deben anonimizar del presente periodo son: `areaa_asignatura`.

En diferente sentido, se realizará el análisis visual de la comparativa del segundo periodo a continuación:

Ilustración 479

Comparativa de tipos de datos del segundo periodo de los datos

Periodo 2021-02

`perdidas[1].dtypes`

<code>estudiante</code>	<code>object</code>
<code>periodo</code>	<code>int64</code>
<code>facultad</code>	<code>object</code>
<code>carrera</code>	<code>object</code>
<code>nivel</code>	<code>object</code>
<code>area_asignatura</code>	<code>int64</code>
<code>aprobado</code>	<code>object</code>
<code>creditos</code>	<code>float64</code>
<code>dtype:</code>	<code>object</code>

`calificaciones[1].dtypes`

<code>estudiante</code>	<code>object</code>
<code>periodo</code>	<code>int64</code>
<code>facultad</code>	<code>object</code>
<code>docente</code>	<code>object</code>
<code>area_asignatura</code>	<code>object</code>
<code>carrera</code>	<code>object</code>
<code>dtype:</code>	<code>object</code>

`becas[1].dtypes`

<code>estudiante</code>	<code>object</code>
<code>etnia</code>	<code>object</code>
<code>nivel</code>	<code>object</code>
<code>segundas_matriculas</code>	<code>int64</code>
<code>terceras_matriculas</code>	<code>int64</code>
<code>dtype:</code>	<code>object</code>

A partir de la comparativa, se puede concluir que la variable `area_asignatura` en el reporte de asignaturas perdidas es de tipo diferente del reporte de calificaciones por carrera, por lo que nuevamente es la única variable con necesidad de armonizarse.

En tercer lugar, se visualizará la comparativa del tercer periodo de los datos para su respectiva evaluación visual:

Ilustración 480

Comparativa de tipos de datos del tercer periodo de los datos

Periodo 2022-01

`perdidas[2].dtypes`

<code>estudiante</code>	<code>object</code>
<code>periodo</code>	<code>int64</code>
<code>facultad</code>	<code>object</code>
<code>carrera</code>	<code>object</code>
<code>nivel</code>	<code>object</code>
<code>area_asignatura</code>	<code>int64</code>
<code>aprobado</code>	<code>object</code>
<code>creditos</code>	<code>float64</code>
<code>dtype:</code>	<code>object</code>

`calificaciones[2].dtypes`

<code>estudiante</code>	<code>object</code>
<code>periodo</code>	<code>int64</code>
<code>facultad</code>	<code>object</code>
<code>docente</code>	<code>object</code>
<code>area_asignatura</code>	<code>object</code>
<code>carrera</code>	<code>object</code>
<code>dtype:</code>	<code>object</code>

`becas[2].dtypes`

<code>estudiante</code>	<code>object</code>
<code>etnia</code>	<code>object</code>
<code>nivel</code>	<code>object</code>
<code>segundas_matriculas</code>	<code>int64</code>
<code>terceras_matriculas</code>	<code>int64</code>
<code>dtype:</code>	<code>object</code>

A partir de la comparativa, se puede concluir que la variable `area_asignatura` en el reporte de asignaturas perdidas es de tipo diferente del reporte de calificaciones por carrera, por lo que,

para el tercer periodo por tercera vez la variable `área_asignatura` es la única variable que requiere de una armonización.

Finalmente, a continuación, se exhibirá la comparativa del cuarto y último periodo de los datos:

Ilustración 481

Comparativa de tipos de datos del cuarto periodo de los datos

Periodo 2022-02

`perdidas[3].dtypes`

<code>estudiante</code>	<code>object</code>
<code>periodo</code>	<code>int64</code>
<code>facultad</code>	<code>object</code>
<code>carrera</code>	<code>object</code>
<code>nivel</code>	<code>float64</code>
<code>area_asignatura</code>	<code>int64</code>
<code>aprobado</code>	<code>object</code>
<code>creditos</code>	<code>float64</code>
<code>dtype: object</code>	

`calificaciones[3].dtypes`

<code>estudiante</code>	<code>object</code>
<code>periodo</code>	<code>int64</code>
<code>facultad</code>	<code>object</code>
<code>docente</code>	<code>object</code>
<code>area_asignatura</code>	<code>object</code>
<code>carrera</code>	<code>object</code>
<code>dtype: object</code>	

`becas[3].dtypes`

<code>estudiante</code>	<code>object</code>
<code>etnia</code>	<code>object</code>
<code>nivel</code>	<code>object</code>
<code>segundas_matriculas</code>	<code>int64</code>
<code>terceras_matriculas</code>	<code>int64</code>
<code>dtype: object</code>	

En esta ocasión, se puede concluir que la variable `area_asignatura` en el reporte de asignaturas perdidas es de tipo diferente del reporte de calificaciones por carrera, y, adicionalmente, se puede concluir que la variable `nivel` en el reporte de asignaturas perdidas es

de tipo diferente del reporte de becas. Esto quiere decir que las variables que requieren de una armonización en este periodo son: `area_asignatura` y `nivel`.

En resumen, tras el análisis realizado, se puede determinar que las siguientes variables necesitan ser armonizadas dado el periodo y reporte:

Tabla 37
Identificación de las variables que requieren armonización

Pe- riodo	Variables	Reportes involucrados
2021-01	<code>area_asignatura</code>	Reporte de Asignaturas perdidas y el reporte de Calificaciones por carrera
2021-02	<code>area_asignatura</code>	Reporte de Asignaturas perdidas y el reporte de Calificaciones por carrera
2022-01	<code>area_asignatura</code>	Reporte de Asignaturas perdidas y el reporte de Calificaciones por carrera
2022-02	<code>area_asignatura</code>	Reporte de Asignaturas perdidas y el reporte de Calificaciones por carrera
2022-02	<code>nivel</code>	Reporte de Asignaturas perdidas y el Becas

Sin embargo, no es prudente realizar un único proceso de cambio de tipo de dato sin antes realizar un análisis de los datos que poseen cada una de estas variables, esto, con el fin de escoger como investigador el tipo de dato más apropiado para la variable en cuestiones de compatibilidad para la unión de datos. El análisis de la variable `area_asignatura` podrá ser visualizado a continuación:

Ilustración 482

Código para llevar a cabo el análisis de la variable area_asignatura

```
periodos = ["2021-01", "2021-02", "2022-01", "2022-02"]

# Se itera por todos los periodos
for i in range(4):
    print(f"\nAnalizando el periodo {periodos[i]}...")

    # Impresión de los valores únicos de la variable en el reporte
    print("Analizando el reporte de asignaturas perdidas...")
    for variable in perdidas[i]["area_asignatura"].unique():
        print("\t-", variable)

    # Impresión de los valores únicos de la variable en el reporte
    print("\nAnalizando el reporte de Calificaciones por carrera...")
    for variable in calificaciones[i]["area_asignatura"].unique():
        print("\t-", variable)
```

Ilustración 483

Resultado del primer periodo del análisis de la variable area_asignatura

```
Analizando el periodo 2021-01...
Analizando el reporte de asignaturas perdidas...
- 9
- 3
- 4
- 2
- 5
- 7
- 6
- 11
- 1

Analizando el reporte de Calificaciones por carrera...
- Cien.Natur.Matemát.y Estadíst.
- Administración
- Artes y Humanidades
- Investigación
- Ingeniería Industria y Constr.
- Salud y Bienestar
- Cien.Soc.period.form y Derecho
- Tecno.Informac.y Comun.(TIC)
- Agricul.Silvic.pesca y veteri.
- Educación
- Servicios
```

Ilustración 484

Resultado del segundo periodo del análisis de la variable area_asignatura

```
Analizando el periodo 2021-02...
Analizando el reporte de asignaturas perdidas...
- 5
- 3
- 4
- 7
- 2
- 11
- 1
- 6
- 10
- 9

Analizando el reporte de Calificaciones por carrera...
- Administración
- Cien.Soc.period.form y Derecho
- Cien.Natur.Matemát.y Estadíst.
- Artes y Humanidades
- Tecno.Informac.y Comun.(TIC)
- Ingeniería Industria y Constr.
- Educación
- Salud y Bienestar
- Investigación
- Servicios
- Agricul.Silvic.pesca y veteri.
```

Ilustración 485

Resultado del tercer periodo del análisis de la variable area_asignatura

```
Analizando el periodo 2022-01...
Analizando el reporte de asignaturas perdidas...
  - 5
  - 7
  - 4
  - 3
  - 6
  - 9
  - 11
  - 2
  - 10
  - 1
  - 8

Analizando el reporte de Calificaciones por carrera...
  - Administración
  - Cien.Natur.Matemát.y Estadíst.
  - Artes y Humanidades
  - Educación
  - Investigación
  - Cien.Soc.period.form y Derecho
  - Ingeniería Industria y Constr.
  - Salud y Bienestar
  - Tecno.Informac.y Comun.(TIC)
  - Agricul.Silvic.pesca y veteri.
  - Servicios
```

Ilustración 486

Resultado del cuarto periodo del análisis de la variable area_asignatura

```
Analizando el periodo 2022-02...
Analizando el reporte de asignaturas perdidas...
- 5
- 3
- 4
- 7
- 6
- 9
- 2
- 11
- 10
- 1
- 8

Analizando el reporte de Calificaciones por carrera...
- Artes y Humanidades
- Investigación
- Salud y Bienestar
- Cien.Natur.Matemát.y Estadíst.
- Administración
- Tecno.Informac.y Comun.(TIC)
- Cien.Soc.period.form y Derecho
- Educación
- Ingeniería Industria y Constr.
- Agricul.Silvic.pesca y veteri.
- Servicios
```

En consideración del análisis realizado, se puede observar que, para todos los periodos, los datos de las áreas de las asignaturas en el reporte de Asignaturas perdidas se encuentran en código mientras que en el reporte de Calificaciones por Carrera no lo están. Por lo que, para armonizar la variable area_asignaturas en todos los periodos, se ha decidido transformar los registros de la variable en el reporte de Calificaciones por carrera a sus respectivos códigos numéricos, dicha elección se basa en que de esta manera se simplificará la complejidad y aumentará la simpleza en el desarrollo de las siguientes etapas de este capítulo.

El código utilizado para la armonización de la variable area_asignatura fue el siguiente:

Ilustración 487

Código para la armonización de la variable area_asignatura

```
# Definición de un diccionario de las áreas de asignaturas con sus códigos
codigos_asignaturas = {
    "Educación" : 1,
    "Artes y Humanidades" : 2,
    "Cien.Soc.period.form y Derecho" : 3,
    "Administración" : 4,
    "Cien.Natur.Matemát.y Estadíst." : 5,
    "Tecno.Informac.y Comun.(TIC)" : 6,
    "Ingeniería Industria y Constr." : 7,
    "Agricul.Silvic.pesca y veteri." : 8,
    "Salud y Bienestar" : 9,
    "Servicios" : 10,
    "Investigación" : 11
}

# Se armoniza el reporte de calificaciones por carrera
for calificacion in calificaciones:
    calificacion["area_asignatura"].replace(codigos_asignaturas, inplace=True)
```

Una vez se concluyó la armonización de la variable area_asignatura en función de la decisión tomada, se puede apreciar una vista previa de la armonización donde, ahora, la variable area_asignatura posee códigos y ya no nombres.

Ilustración 488

Vista previa de la armonización de la variable area_asignatura

```
# Vista previa de la armonización
calificaciones[0].head()
```

	estudiante	periodo	facultad	docente	area_asignatura	carrera
0	E8695	202161	Facultad Cs Adm. y Contables	D936	5	Administración de Empresas
4	E8695	202161	Facultad Cs Adm. y Contables	D619	4	Administración de Empresas
8	E8695	202161	Facultad Cs Adm. y Contables	D303	2	Administración de Empresas
12	E8695	202161	Facultad Cs Adm. y Contables	D269	4	Administración de Empresas
16	E8695	202161	Facultad Cs Exactas Naturales	D782	11	Administración de Empresas

Sin embargo, el preprocesamiento de la unión aun no finaliza, puesto que es necesario realizar la armonización de la variable nivel, para ello se requiere de un análisis que puede ser observado a continuación:

Ilustración 489

Código para llevar a cabo el análisis de la variable nivel

```
periodos = ["2021-01", "2021-02", "2022-01", "2022-02"]

# Se itera por todos los periodos
for i in range(4):
    print(f"\nAnalizando el periodo {periodos[i]}...")

    # Impresión de los valores únicos de la variable en el reporte
    print("Analizando el reporte de asignaturas perdidas...")
    for variable in perdidas[i]["nivel"].unique():
        print("\t-", variable)

    # Impresión de los valores únicos de la variable en el reporte
    print("\nAnalizando el reporte de Becas...")
    for variable in becas[i]["nivel"].unique():
        print("\t-", variable)
```

Ilustración 490

Resultado del primer periodo del análisis de la variable nivel

```
Analizando el periodo 2021-01...
Analizando el reporte de asignaturas perdidas...
    - 09
    - 04
    - 03
    - 08
    - 05
    - 01
    - 07
    - 02
    - 06
    - 10
    - Electiva
    - nan

Analizando el reporte de Becas...
    - 01
    - 07
    - 04
    - 10
    - 02
    - 06
    - 05
    - 03
    - nan
    - 08
    - 09
    - EL
    - ELECT
```

Ilustración 491

Resultado del segundo periodo del análisis de la variable nivel

```
Analizando el periodo 2021-02...
Analizando el reporte de asignaturas perdidas...
- 07
- 04
- 06
- 05
- 09
- 03
- 08
- 01
- 10
- 02
- nan
- Electiva

Analizando el reporte de Becas...
- 08
- 05
- 03
- 07
- 01
- 02
- 06
- 04
- 09
- 10
- nan
- 89
- EL
- ELECT
```

Ilustración 492

Resultado del tercer periodo del análisis de la variable nivel

```
Analizando el periodo 2022-01...
Analizando el reporte de asignaturas perdidas...
- 02
- 08
- 04
- 06
- 07
- 05
- 09
- nan
- 01
- 03
- 10
- Electiva

Analizando el reporte de Becas...
- 06
- 01
- 07
- 08
- 02
- 04
- 03
- 10
- 09
- 05
- nan
- EL
- 89
- ELEC
```

Ilustración 493

Resultado del cuarto periodo del análisis de la variable nivel

```
Analizando el periodo 2022-02...
Analizando el reporte de asignaturas perdidas...
- 3.0
- 7.0
- 8.0
- 9.0
- 1.0
- 6.0
- 4.0
- 5.0
- nan
- 10.0
- 2.0

Analizando el reporte de Becas...
- 07
- 02
- 03
- 05
- 04
- 08
- 01
- 89
- 09
- 06
- nan
- 10
- EL
```

Tomando en cuenta el análisis llevado a cabo sobre la variable nivel, se puede observar que los datos en el reporte de Asignaturas perdidas del cuarto periodo se encuentran como números de punto flotante mientras que los del reporte de Becas son de tipo "object", así como los demás periodos del reporte de asignaturas perdidas; por lo tanto, se ha tomado la decisión de, dado que, armonizará la variable nivel del cuarto periodo del reporte de Asignaturas perdidas para que sea de tipo object con la misma estructura visible en los datos de la variable del reporte de Becas. Para esto, es necesario realizar el siguiente código:

Ilustración 494

Código de armonización de la variable nivel del cuarto periodo del reporte de Asignaturas Perdidas

```
# Se realizan los cambios para ajustar lo máximo posible al formato presentado en el análisis anterior
perdidas[3]["nivel"] = perdidas[3]["nivel"].apply(lambda x: str(x).split(".")[0] if str(x) != "nan" else x)
perdidas[3]["nivel"] = perdidas[3]["nivel"].apply(lambda x: "0" + str(x) if len(str(x)) < 2 else str(x))
```

El código presente ajusta los contenidos de la variable del cuarto periodo del reporte de tal manera que se ajuste de manera idéntica a los demás periodos y reportes, una visualización previa de la armonización realizada se puede admirar a continuación:

Ilustración 495

Vista previa de la armonización de la variable nivel

```
# Vista previa del cambio realizado
perdidas[3]["nivel"].unique()

array(['03', '07', '08', '09', '01', '06', '04', '05', 'nan', '10', '02'],
      dtype=object)
```

Una vez finalizado el proceso de armonización de las variables consideradas incompatibles, es necesario realizar un análisis para verificar que no existan más incompatibilidades en cuestiones de tipos de datos, sin embargo, cabe recalcar que en esta ocasión el análisis será automatizado. El código y sus resultados se pueden admirar a continuación:

Ilustración 496

Análisis de la compatibilidad entre variables armonizadas

```
# Periodos existentes
periodos = ["2021-01", "2021-02", "2022-01", "2022-02"]

# Variables que se relacionan
perdidas_calificaciones = ["estudiante", "periodo", "facultad", "carrera", "area_asignatura"]
perdidas_becas = ["estudiante", "nivel"]
calificaciones_becas = ["estudiante"]

# Se itera en todos los periodos
for i in range(4):
    print(f"\nAnalizando el periodo {periodos[i]}...")

    # Verificación del reporte de asignaturas perdidas y calificaciones por carrera
    print(f"\n\tAnalizando las variables que se relacionan entre el reporte {perdidas[i].name} y {calificaciones[i].name}...")
    for variable in perdidas_calificaciones:
        print(f"\t\t{variable}: {perdidas[i][variable].dtype == calificaciones[i][variable].dtype}")

    # Verificación del reporte de asignaturas perdidas y becas
    print(f"\n\tAnalizando las variables que se relacionan entre el reporte {perdidas[i].name} y {becas[i].name}...")
    for variable in perdidas_becas:
        print(f"\t\t{variable}: {perdidas[i][variable].dtype == becas[i][variable].dtype}")

    # Verificación del reporte de calificaciones por carrera y becas
    print(f"\n\tAnalizando las variables que se relacionan entre el reporte {calificaciones[i].name} y {becas[i].name}...")
    for variable in calificaciones_becas:
        print(f"\t\t{variable}: {becas[i][variable].dtype == calificaciones[i][variable].dtype}")

    print("\n" + "*" * 100)
```

Ilustración 497

Resultado del primer periodo del análisis de compatibilidad

```
Analizando el periodo 2021-01...

Analizando las variables que se relacionan entre el reporte perdidas01 y calificaciones01...
    estudiante: True
    periodo: True
    facultad: True
    carrera: True
    area_asignatura: True

Analizando las variables que se relacionan entre el reporte perdidas01 y becas01...
    estudiante: True
    nivel: True

Analizando las variables que se relacionan entre el reporte calificaciones01 y becas01...
    estudiante: True
```

Ilustración 498

Resultado del segundo periodo del análisis de compatibilidad

```
Analizando el periodo 2021-02...

Analizando las variables que se relacionan entre el reporte perdidas02 y calificaciones02...
  estudiante: True
  periodo: True
  facultad: True
  carrera: True
  area_asignatura: True

Analizando las variables que se relacionan entre el reporte perdidas02 y becas02...
  estudiante: True
  nivel: True

Analizando las variables que se relacionan entre el reporte calificaciones02 y becas02...
  estudiante: True
```

Ilustración 499

Resultado del tercer periodo del análisis de compatibilidad

```
Analizando el periodo 2022-01...

Analizando las variables que se relacionan entre el reporte perdidas03 y calificaciones03...
  estudiante: True
  periodo: True
  facultad: True
  carrera: True
  area_asignatura: True

Analizando las variables que se relacionan entre el reporte perdidas03 y becas03...
  estudiante: True
  nivel: True

Analizando las variables que se relacionan entre el reporte calificaciones03 y becas03...
  estudiante: True
```

Ilustración 500

Resultado del cuarto periodo del análisis de compatibilidad

```
Analizando el periodo 2022-02...

Analizando las variables que se relacionan entre el reporte perdidas04 y calificaciones04...
  estudiante: True
  periodo: True
  facultad: True
  carrera: True
  area_asignatura: True

Analizando las variables que se relacionan entre el reporte perdidas04 y becas04...
  estudiante: True
  nivel: True

Analizando las variables que se relacionan entre el reporte calificaciones04 y becas04...
  estudiante: True
```

En síntesis, a partir de este momento del trabajo de titulación, luego de realizar los análisis y armonizaciones pertinentes, se puede concluir que los conjuntos de datos

correspondientes a los reportes de Asignaturas perdidas, Calificaciones por carrera y Becas están listos para ser unificados independiente del periodo escogido.

Unificación de los conjuntos de datos por cada periodo

En esta sección se realizará la unificación de los conjuntos de datos dado el periodo, pero, es necesario recalcar que no se realizará la unificación entre los periodos del mismo reporte por el momento, en otras palabras, el objetivo de esta sección es unir los conjuntos de datos de cada reporte dado el periodo, es decir, los conjuntos de datos del periodo 2021-01, después los del 2021-02, etc. No obstante, es necesario mencionar que la manera en la que se realizaran las uniones será en un primer lugar la unión de reporte de Calificaciones por carrera y de Becas debido a que ambas presentan la relación más pequeña, después se realizará la unión con el reporte de Asignaturas perdidas. Esto se debe a los siguientes puntos:

Minimización de la complejidad. Unir conjuntos de datos con muchas relaciones puede ser más complejo y susceptible a errores, por lo que, al comenzar con los conjuntos de datos con menos relaciones, se puede reducir la probabilidad de cometer errores y facilitar la depuración del código.

Control de los datos faltantes. Al realizar una unión 'outer', pueden surgir valores faltantes (NaN) cuando no hay correspondencia entre las tablas, a razón de ello, al unir primero las tablas con menos relaciones, se puede tener un mejor control y visibilidad sobre dónde y por qué están surgiendo estos valores faltantes, lo que puede ayudar a decidir cómo manejarlos.

Tomando en cuenta las consideraciones iniciales, se comenzará realizando la unión del primer periodo del reporte de calificaciones por carrera con el reporte becas:

Ilustración 501

Unión del primer periodo reporte de Calificaciones por Carrera y el reporte de Becas

```
# Se seleccionan las variables que se relacionan entre reportes
variables = ["estudiante"]

# Se realiza una primera unión entre el reporte de becas y de calificaciones
union_temp = pd.merge(becas[0], calificaciones[0], on=variables, how="outer")
```

Una vez realizada la unión, es necesario evaluarla en busca de posibles errores, para esto, se ha realizado una evaluación visual general y específica, las mismas pueden observarse a continuación:

Ilustración 502

Evaluación del estado de la unión

union_temp.head(10)										
	estudiante	etnia	nivel	segundas_matriculas	terceras_matriculas	periodo	facultad	docente	area_asignatura	carrera
0	E9532	NaN	01	1	0	202161.0	Dirección Pastoral Univer	D776	2.0	Enfermería
1	E9532	NaN	01	1	0	202161.0	Facultad Cs Exactas Naturales	D699	5.0	Enfermería
2	E9532	NaN	01	1	0	202161.0	Facultad Enfermería	D141	9.0	Enfermería
3	E9532	NaN	01	1	0	202161.0	Facultad Enfermería	D152	3.0	Enfermería
4	E2371	Mestizo	07	0	0	202161.0	Facultad Economía	D177	3.0	Economía
5	E2371	Mestizo	07	0	0	202161.0	Facultad Economía	D172	3.0	Economía
6	E2371	Mestizo	07	0	0	202161.0	Facultad Economía	D158	3.0	Economía
7	E2371	Mestizo	07	0	0	202161.0	Facultad Economía	D579	3.0	Economía
8	E2371	Mestizo	07	0	0	202161.0	Facultad Economía	D507	3.0	Economía
9	E2371	Mestizo	07	0	0	202161.0	Facultad Economía	D955	3.0	Economía

Ilustración 503

Evaluación específica de la unión

```
print(f"La unión posee {union_temp.shape[0]} registros y {union_temp.shape[1]} variables.")
La unión posee 40885 registros y 10 variables.

print(f"El reporte de calificaciones por carrera posee {calificaciones[0].shape[0]} registros y {calificaciones[0].shape[1]} variables.")
print(f"El reporte de becas posee {becas[0].shape[0]} registros y {becas[0].shape[1]} variables.")

El reporte de calificaciones por carrera posee 39861 registros y 6 variables.
El reporte de becas posee 8608 registros y 5 variables.
```

A partir de las evaluaciones realizadas, se puede determinar que el resultado de la unión inicial se encuentra dentro de un rango aceptable, sin embargo, se aprecia cómo probablemente existen algunos registros del reporte de becas que no existían en el reporte de calificaciones finales o viceversa. De todas maneras, es factible continuar con la unión del primer periodo del

reporte de asignaturas perdidas, el código y las evaluaciones de este se pueden encontrar a continuación:

Ilustración 504

Unión del primer periodo del reporte de asignaturas perdidas con la unión inicial

```
# Se seleccionan las variables que se relacionan entre la unión ya realizada y el reporte de asignaturas perdidas
variables = ["estudiante", "periodo", "nivel", "facultad", "carrera", "area_asignatura"]

# Se realiza la unión de todos los reportes del periodo 2021-01
periodo01 = pd.merge(union_temp, perdidas[0], on=variables, how="outer")
```

Ilustración 505

Evaluación del estado de la unión completa

periodo01.head(10)

	estudiante	etnia	nivel	segundas_matriculas	terceras_matriculas	periodo	facultad	docente	area_asignatura	carrera	aprobado	creditos
0	E9532	NaN	01	1.0	0.0	202161.0	Dirección Pastoral Univer	D776	2.0	Enfermería	NaN	NaN
1	E9532	NaN	01	1.0	0.0	202161.0	Facultad Cs Exactas Naturales	D699	5.0	Enfermería	NaN	NaN
2	E9532	NaN	01	1.0	0.0	202161.0	Facultad Enfermería	D141	9.0	Enfermería	NaN	NaN
3	E9532	NaN	01	1.0	0.0	202161.0	Facultad Enfermería	D152	3.0	Enfermería	NaN	NaN
4	E2371	Mestizo	07	0.0	0.0	202161.0	Facultad Economía	D177	3.0	Economía	NaN	NaN
5	E2371	Mestizo	07	0.0	0.0	202161.0	Facultad Economía	D172	3.0	Economía	NaN	NaN
6	E2371	Mestizo	07	0.0	0.0	202161.0	Facultad Economía	D158	3.0	Economía	NaN	NaN
7	E2371	Mestizo	07	0.0	0.0	202161.0	Facultad Economía	D579	3.0	Economía	NaN	NaN
8	E2371	Mestizo	07	0.0	0.0	202161.0	Facultad Economía	D507	3.0	Economía	NaN	NaN
9	E2371	Mestizo	07	0.0	0.0	202161.0	Facultad Economía	D955	3.0	Economía	NaN	NaN

Ilustración 506

Evaluación específica de la unión completa

```
print(f"La unión posee {periodo01.shape[0]} registros y {periodo01.shape[1]} variables.")
La unión posee 41936 registros y 12 variables.

print(f"La unión inicial poseía {union_temp.shape[0]} registros y {union_temp.shape[1]} variables.")
La unión inicial poseía 40885 registros y 10 variables.

print(f"El reporte de calificaciones por carrera posee {perdidas[0].shape[0]} registros y {perdidas[0].shape[1]} variables.")
print(f"El reporte de calificaciones por carrera posee {calificaciones[0].shape[0]} registros y {calificaciones[0].shape[1]} variables.")
print(f"El reporte de becas posee {becas[0].shape[0]} registros y {becas[0].shape[1]} variables.")

El reporte de calificaciones por carrera posee 1292 registros y 8 variables.
El reporte de calificaciones por carrera posee 39861 registros y 6 variables.
El reporte de becas posee 8608 registros y 5 variables.
```

Se puede observar que nuevamente la unión se encuentra dentro de un rango aceptable, pero también es posible que algunos registros de algún reporte no se encuentren en los demás. Por lo que, de manera concluyente, se puede concretar que a pesar de que se presenten algunos registros adicionales a los aproximados teóricamente para la unión, el resultado de esta se encuentra dentro de un rango real y posible, en otras palabras, la unión fue un éxito.

Por otro lado, tomando en cuenta el éxito de la unión de los primeros periodos de los reportes, se procederá a realizar la unión de los segundos periodos comenzando por los reportes de calificaciones por carrera y becas, el código y las evaluaciones se pueden encontrar a continuación:

Ilustración 507

Unión del segundo periodo reporte de Calificaciones por Carrera y el reporte de Becas

```
: # Se seleccionan las variables que se relacionan entre reportes
variables = ["estudiante"]

# Se realiza una primera unión entre el reporte de becas y de calificaciones
union_temp = pd.merge(becas[1], calificaciones[1], on=variables, how="outer")
```

Ilustración 508

Evaluación del estado de la unión

union_temp.head(10)										
	estudiante	etnia	nivel	segundas_matriculas	terceras_matriculas	periodo	facultad	docente	area_asignatura	carrera
0	E2371	Mestizo	08	0	0	202166.0	Facultad Economía	D579	3.0	Economía
1	E2371	Mestizo	08	0	0	202166.0	Facultad Economía	D126	3.0	Economía
2	E2371	Mestizo	08	0	0	202166.0	Facultad Economía	D682	3.0	Economía
3	E2371	Mestizo	08	0	0	202166.0	Facultad Economía	D639	3.0	Economía
4	E2371	Mestizo	08	0	0	202166.0	Facultad Economía	D728	3.0	Economía
5	E2838	NaN	05	0	0	202166.0	Facultad Cs Adm. y Contables	D452	4.0	Negocios Internacionales
6	E2838	NaN	05	0	0	202166.0	Facultad Cs Adm. y Contables	D106	4.0	Negocios Internacionales
7	E2838	NaN	05	0	0	202166.0	Facultad Cs Adm. y Contables	D738	4.0	Negocios Internacionales
8	E2838	NaN	05	0	0	202166.0	Facultad Cs Adm. y Contables	D620	3.0	Negocios Internacionales
9	E2838	NaN	05	0	0	202166.0	Facultad Cs Adm. y Contables	D48	4.0	Negocios Internacionales

Ilustración 509

Evaluación específica de la unión

```
print(f"La unión posee {union_temp.shape[0]} registros y {union_temp.shape[1]} variables.")
La unión posee 39940 registros y 10 variables.

print(f"El reporte de calificaciones por carrera posee {calificaciones[1].shape[0]} registros y {calificaciones[1].shape[1]} variables.")
print(f"El reporte de becas posee {becas[1].shape[0]} registros y {becas[1].shape[1]} variables.")
El reporte de calificaciones por carrera posee 39001 registros y 6 variables.
El reporte de becas posee 8562 registros y 5 variables.
```

De la unión, se puede concluir que el resultado de la unión inicial se encuentra dentro de un rango aceptable, sin embargo, se aprecia cómo probablemente existen algunos registros del reporte de becas que no existían en el reporte de calificaciones finales o viceversa, pero en menor

cantidad. De todas maneras, es adecuado para continuar con la unión del segundo periodo del reporte de asignaturas perdidas, el código y evaluaciones se pueden encontrar a continuación:

Ilustración 510

Unión del segundo periodo del reporte de asignaturas perdidas con la unión inicial

```
# Se seleccionan las variables que se relacionan entre la unión ya realizada y el reporte de asignaturas perdidas
variables = ["estudiante", "periodo", "nivel", "facultad", "carrera", "area_asignatura"]

# Se realiza la unión de todos los reportes del periodo 2021-01
periodo02 = pd.merge(union_temp, perdidas[1], on=variables, how="outer")
```

Ilustración 511

Evaluación del estado de la unión completa

periodo01.head(10)

	estudiante	etnia	nivel	segundas_matriculas	terceras_matriculas	periodo	facultad	docente	area_asignatura	carrera	aprobado	creditos
0	E9532	NaN	01	1.0	0.0	202161.0	Dirección Pastoral Univer	D776	2.0	Enfermería	NaN	NaN
1	E9532	NaN	01	1.0	0.0	202161.0	Facultad Cs Exactas Naturales	D699	5.0	Enfermería	NaN	NaN
2	E9532	NaN	01	1.0	0.0	202161.0	Facultad Enfermería	D141	9.0	Enfermería	NaN	NaN
3	E9532	NaN	01	1.0	0.0	202161.0	Facultad Enfermería	D152	3.0	Enfermería	NaN	NaN
4	E2371	Mestizo	07	0.0	0.0	202161.0	Facultad Economía	D177	3.0	Economía	NaN	NaN
5	E2371	Mestizo	07	0.0	0.0	202161.0	Facultad Economía	D172	3.0	Economía	NaN	NaN
6	E2371	Mestizo	07	0.0	0.0	202161.0	Facultad Economía	D158	3.0	Economía	NaN	NaN
7	E2371	Mestizo	07	0.0	0.0	202161.0	Facultad Economía	D579	3.0	Economía	NaN	NaN
8	E2371	Mestizo	07	0.0	0.0	202161.0	Facultad Economía	D507	3.0	Economía	NaN	NaN
9	E2371	Mestizo	07	0.0	0.0	202161.0	Facultad Economía	D955	3.0	Economía	NaN	NaN

Ilustración 512

Evaluación específica de la unión completa

```
print(f"La unión posee {periodo01.shape[0]} registros y {periodo01.shape[1]} variables.")
La unión posee 41936 registros y 12 variables.

print(f"La unión inicial poseía {union_temp.shape[0]} registros y {union_temp.shape[1]} variables.")
La unión inicial poseía 39940 registros y 10 variables.

print(f"El reporte de calificaciones por carrera posee {perdidas[1].shape[0]} registros y {perdidas[1].shape[1]} variables.")
print(f"El reporte de calificaciones por carrera posee {calificaciones[1].shape[0]} registros y {calificaciones[1].shape[1]} variables.")
print(f"El reporte de becas posee {becas[1].shape[0]} registros y {becas[1].shape[1]} variables.")

El reporte de calificaciones por carrera posee 1668 registros y 8 variables.
El reporte de calificaciones por carrera posee 39001 registros y 6 variables.
El reporte de becas posee 8562 registros y 5 variables.
```

De la unión completa se puede observar que nuevamente la unión se encuentra dentro de un rango aceptable, pero también es posible que algunos registros de algún reporte no se encuentren en los demás. Por lo que, se puede concluir que a pesar de que se presenten algunos registros adicionales a los aproximados teóricamente para la unión, el resultado de esta se encuentra dentro de un rango real y posible.

En tercer orden, se presentará a continuación la unión de los terceros periodos de los reportes, inicialmente de nuevo por la unión del reporte de calificaciones por carrera y el reporte de becas, el código y evaluaciones se pueden encontrar a continuación:

Ilustración 513

Unión del tercero periodo reporte de Calificaciones por Carrera y el reporte de Becas

```
# Se seleccionan las variables que se relacionan entre reportes
variables = ["estudiante"]

# Se realiza una primera unión entre el reporte de becas y de calificaciones
union_temp = pd.merge(becas[2], calificaciones[2], on=variables, how="outer")
```

Ilustración 514

Evaluación del estado de la unión

union_temp.head(10)											
	estudiante	etnia	nivel	segundas_matriculas	terceras_matriculas	periodo	facultad	docente	area_asignatura	carrera	
0	E2838	NaN	06	0	0	202261.0	Facultad Cs Adm. y Contables	D199	4.0	Negocios Internacionales	
1	E2838	NaN	06	0	0	202261.0	Facultad Cs Adm. y Contables	D784	3.0	Negocios Internacionales	
2	E2838	NaN	06	0	0	202261.0	Facultad Cs Adm. y Contables	D267	4.0	Negocios Internacionales	
3	E2838	NaN	06	0	0	202261.0	Facultad Cs Adm. y Contables	D213	4.0	Negocios Internacionales	
4	E2838	NaN	06	0	0	202261.0	Facultad Cs Adm. y Contables	D758	4.0	Negocios Internacionales	
5	E2838	NaN	06	0	0	202261.0	Facultad Cs Adm. y Contables	D439	4.0	Negocios Internacionales	
6	E5194	NaN	01	0	0	202261.0	Facultad Economía	D955	3.0	Economía	
7	E5194	NaN	01	0	0	202261.0	Facultad Economía	D2	3.0	Economía	
8	E5194	NaN	01	0	0	202261.0	Facultad Economía	D289	4.0	Economía	
9	E5194	NaN	01	0	0	202261.0	Facultad Economía	D815	3.0	Economía	

Ilustración 515

Evaluación específica de la unión

```
print(f"La unión posee {union_temp.shape[0]} registros y {union_temp.shape[1]} variables.")
La unión posee 37756 registros y 10 variables.

print(f"El reporte de calificaciones por carrera posee {calificaciones[2].shape[0]} registros y {calificaciones[2].shape[1]} variables.")
print(f"El reporte de becas posee {becas[2].shape[0]} registros y {becas[2].shape[1]} variables.")
El reporte de calificaciones por carrera posee 36703 registros y 6 variables.
El reporte de becas posee 8240 registros y 5 variables.
```

Dada la evaluación realizada, se puede observar que el resultado de la unión inicial se encuentra dentro de un rango aceptable, sin embargo, se aprecia cómo probablemente existen algunos registros del reporte de becas que no existían en el reporte de calificaciones finales o viceversa, por otra parte, se puede visualizar una cantidad mayor de valores nulos en la variable

de etnia, esto resalta la importancia de realizar un análisis sobre la razón de esta ocurrencia; dicho análisis con su código se pueden encontrar a continuación:

Ilustración 516

Análisis de la variable etnia

```
print(f"Existen {becas[2].shape[0]} registros el reporte de becas del periodo 2022-01.")

# Almacena la cantidad de registros faltantes
cantidad = becas[2]["etnia"].isna().sum()

# Se presenta la cantidad de valores faltantes
print(f"La cantidad de registros faltantes en la variable 'etnia' del reporte de becas es: {cantidad}")

# Se presenta el porcentaje de valores faltantes del total
print(f"El porcentaje de valores faltantes de la variable 'etnia' del reporte de becas es: {round(cantidad / becas[2].shape[0] * 100, 2)}%")

# Se presentan los datos de valores faltantes en la unión
cantidad = union_temp["etnia"].isna().sum()
print(f"La cantidad de registros faltantes en la variable 'etnia' del reporte de becas es: {cantidad}")
print(f"El porcentaje de valores faltantes de la variable 'etnia' del reporte de becas es: {round(cantidad / union_temp.shape[0] * 100, 2)}%")

Existen 8240 registros el reporte de becas del periodo 2022-01.
La cantidad de registros faltantes en la variable 'etnia' del reporte de becas es: 5952
El porcentaje de valores faltantes de la variable 'etnia' del reporte de becas es: 72.23%
La cantidad de registros faltantes en la variable 'etnia' del reporte de becas es: 28356
El porcentaje de valores faltantes de la variable 'etnia' del reporte de becas es: 75.1%
```

Dado que el porcentaje de valores nulos de la variable etnia es del 72.23%, se puede comprender cómo en el conjunto de datos producto de la unión se encuentra con un porcentaje de valores faltantes del 75.1%, es decir, a pesar de la visibilidad de la enorme cantidad de valores faltantes, la unión se puede considerar como un éxito. En otras palabras, se puede continuar con la unión del tercer periodo del reporte de becas, el código y las evaluaciones se pueden visualizar a continuación:

Ilustración 517

Unión del tercer periodo del reporte de asignaturas perdidas con la unión inicial

```
: # Se seleccionan las variables que se relacionan entre la unión ya realizada y el reporte de asignaturas perdidas
variables = ["estudiante", "periodo", "nivel", "facultad", "carrera", "area_asignatura"]

# Se realiza la unión de todos los reportes del periodo 2021-01
periodo03 = pd.merge(union_temp, perdidas[2], on=variables, how="outer")
```

Ilustración 518

Evaluación del estado de la unión completa

periodo03.head(10)												
	estudiante	etnia	nivel	segundas_matriculas	terceras_matriculas	periodo	facultad	docente	area_asignatura	carrera	aprobado	creditos
0	E2838	NaN	06	0.0	0.0	202261.0	Facultad Cs Adm. y Contables	D199	4.0	Negocios Internacionales	NaN	NaN
1	E2838	NaN	06	0.0	0.0	202261.0	Facultad Cs Adm. y Contables	D267	4.0	Negocios Internacionales	NaN	NaN
2	E2838	NaN	06	0.0	0.0	202261.0	Facultad Cs Adm. y Contables	D213	4.0	Negocios Internacionales	NaN	NaN
3	E2838	NaN	06	0.0	0.0	202261.0	Facultad Cs Adm. y Contables	D758	4.0	Negocios Internacionales	NaN	NaN
4	E2838	NaN	06	0.0	0.0	202261.0	Facultad Cs Adm. y Contables	D439	4.0	Negocios Internacionales	NaN	NaN
5	E2838	NaN	06	0.0	0.0	202261.0	Facultad Cs Adm. y Contables	D784	3.0	Negocios Internacionales	NaN	NaN
6	E5194	NaN	01	0.0	0.0	202261.0	Facultad Economía	D955	3.0	Economía	NaN	NaN
7	E5194	NaN	01	0.0	0.0	202261.0	Facultad Economía	D2	3.0	Economía	NaN	NaN
8	E5194	NaN	01	0.0	0.0	202261.0	Facultad Economía	D815	3.0	Economía	NaN	NaN
9	E5194	NaN	01	0.0	0.0	202261.0	Facultad Economía	D289	4.0	Economía	NaN	NaN

Ilustración 519

Evaluación específica de la unión completa

```
print(f"La unión posee {periodo01.shape[0]} registros y {periodo01.shape[1]} variables.")
La unión posee 41936 registros y 12 variables.

print(f"La unión inicial poseía {union_temp.shape[0]} registros y {union_temp.shape[1]} variables.")
La unión inicial poseía 37756 registros y 10 variables.

print(f"El reporte de calificaciones por carrera posee {perdidas[2].shape[0]} registros y {perdidas[2].shape[1]} variables.")
print(f"El reporte de calificaciones por carrera posee {calificaciones[2].shape[0]} registros y {calificaciones[2].shape[1]} variables.")
print(f"El reporte de becas posee {becas[2].shape[0]} registros y {becas[2].shape[1]} variables.")

El reporte de calificaciones por carrera posee 2518 registros y 8 variables.
El reporte de calificaciones por carrera posee 36703 registros y 6 variables.
El reporte de becas posee 8240 registros y 5 variables.
```

Se puede observar que nuevamente la unión se encuentra dentro de un rango aceptable, pero también es posible que algunos registros de algún reporte no se encuentren en los demás, en otras palabras, se puede concluir que a pesar de que se presenten algunos registros adicionales a los aproximados teóricamente para la unión, el resultado de esta se encuentra dentro de un rango real y posible.

Finalmente, es momento de realizar la unión de los últimos periodos, comenzando por los reportes de calificaciones finales y de becas:

Ilustración 520

Unión del cuarto periodo reporte de Calificaciones por Carrera y el reporte de Becas

```
# Se seleccionan las variables que se relacionan entre reportes
variables = ["estudiante"]

# Se realiza una primera unión entre el reporte de becas y de calificaciones
union_temp = pd.merge(becas[3], calificaciones[3], on=variables, how="outer")
```

Ilustración 521

Evaluación del estado de la unión

union_temp.head(10)											
	estudiante	etnia	nivel	segundas_matriculas	terceras_matriculas	periodo	facultad	docente	area_asignatura	carrera	
0	E2838	NaN	07	0	0	202266.0	Facultad Cs Adm. y Contables	D884	4.0	Negocios Internacionales	
1	E2838	NaN	07	0	0	202266.0	Facultad Cs Adm. y Contables	D43	4.0	Negocios Internacionales	
2	E2838	NaN	07	0	0	202266.0	Facultad Cs Adm. y Contables	D73	4.0	Negocios Internacionales	
3	E2838	NaN	07	0	0	202266.0	Facultad Cs Adm. y Contables	D153	4.0	Negocios Internacionales	
4	E2838	NaN	07	0	0	202266.0	Facultad Cs Adm. y Contables	D104	4.0	Negocios Internacionales	
5	E2838	NaN	07	0	0	202266.0	Facultad Cs Adm. y Contables	D203	3.0	Negocios Internacionales	
6	E5194	NaN	02	0	0	202266.0	Facultad Economía	D317	3.0	Economía	
7	E5194	NaN	02	0	0	202266.0	Facultad Economía	D295	3.0	Economía	
8	E5194	NaN	02	0	0	202266.0	Facultad Economía	D553	4.0	Economía	
9	E5194	NaN	02	0	0	202266.0	Facultad Economía	D210	3.0	Economía	

Ilustración 522

Evaluación específica de la unión

```
print(f"La unión posee {union_temp.shape[0]} registros y {union_temp.shape[1]} variables.")
La unión posee 36361 registros y 10 variables.

print(f"El reporte de calificaciones por carrera posee {calificaciones[3].shape[0]} registros y {calificaciones[3].shape[1]} variables.")
print(f"El reporte de becas posee {becas[3].shape[0]} registros y {becas[3].shape[1]} variables.")
El reporte de calificaciones por carrera posee 35056 registros y 6 variables.
El reporte de becas posee 8255 registros y 5 variables.
```

De igual forma que para los terceros periodos, se puede observar que el resultado de la unión inicial se encuentra dentro de un rango aceptable, sin embargo, se aprecia cómo probablemente existen algunos registros del reporte de becas que no existían en el reporte de calificaciones finales o viceversa; y, se puede visualizar una cantidad mayor de valores nulos en la variable de etnia.

Ilustración 523

Análisis de la variable etnia

```
print(f'Existen {becas[3].shape[0]} registros el reporte de becas del periodo 2022-01.')

# Almacena la cantidad de registros faltantes
cantidad = becas[3]["etnia"].isna().sum()

# Se presenta la cantidad de valores faltantes
print(f'La cantidad de registros faltantes en la variable 'etnia' del reporte de becas es: {cantidad}')

# Se presenta el porcentaje de valores faltantes del total
print(f'El porcentaje de valores faltantes de la variable 'etnia' del reporte de becas es: {round(cantidad / becas[2].shape[0] * 100, 2)}%')

# Se presentan los datos de valores faltantes en la unión
cantidad = union_temp["etnia"].isna().sum()
print(f'La cantidad de registros faltantes en la variable 'etnia' del reporte de becas es: {cantidad}')
print(f'El porcentaje de valores faltantes de la variable 'etnia' del reporte de becas es: {round(cantidad / union_temp.shape[0] * 100, 2)}%')

Existen 8255 registros el reporte de becas del periodo 2022-01.
La cantidad de registros faltantes en la variable 'etnia' del reporte de becas es: 6818
El porcentaje de valores faltantes de la variable 'etnia' del reporte de becas es: 82.74%
La cantidad de registros faltantes en la variable 'etnia' del reporte de becas es: 31256
El porcentaje de valores faltantes de la variable 'etnia' del reporte de becas es: 85.96%
```

Sin embargo, dado que el porcentaje de valores nulos de la variable etnia es del 82.74%, se puede comprender cómo en el conjunto de datos producto de la unión se encuentra con un porcentaje de valores faltantes del 85.96%, es decir, a pesar de la visibilidad de la enorme cantidad de valores faltantes, la unión se puede considerar como un éxito. Se puede continuar con la unión del tercer periodo del reporte de asignaturas perdidas:

Ilustración 524

Unión del cuarto periodo del reporte de asignaturas perdidas con la unión inicial

```
# Se seleccionan las variables que se relacionan entre la unión ya realizada y el reporte de asignaturas perdidas
variables = ["estudiante", "periodo", "nivel", "facultad", "carrera", "area_asignatura"]

# Se realiza la unión de todos los reportes del periodo 2021-01
periodo04 = pd.merge(union_temp, perdidas[3], on=variables, how="outer")
```

Ilustración 525

Evaluación del estado de la unión completa

periodo04.head(10)													
	estudiante	etnia	nivel	segundas_matriculas	terceras_matriculas	periodo	facultad	docente	area_asignatura	carrera	aprobado	creditos	
0	E2838	NaN	07	0.0	0.0	202266.0	Facultad Cs Adm. y Contables	D884	4.0	Negocios Internacionales	N	0.0	
1	E2838	NaN	07	0.0	0.0	202266.0	Facultad Cs Adm. y Contables	D43	4.0	Negocios Internacionales	N	0.0	
2	E2838	NaN	07	0.0	0.0	202266.0	Facultad Cs Adm. y Contables	D73	4.0	Negocios Internacionales	N	0.0	
3	E2838	NaN	07	0.0	0.0	202266.0	Facultad Cs Adm. y Contables	D153	4.0	Negocios Internacionales	N	0.0	
4	E2838	NaN	07	0.0	0.0	202266.0	Facultad Cs Adm. y Contables	D104	4.0	Negocios Internacionales	N	0.0	
5	E2838	NaN	07	0.0	0.0	202266.0	Facultad Cs Adm. y Contables	D203	3.0	Negocios Internacionales	NaN	NaN	
6	E5194	NaN	02	0.0	0.0	202266.0	Facultad Economía	D317	3.0	Economía	NaN	NaN	
7	E5194	NaN	02	0.0	0.0	202266.0	Facultad Economía	D295	3.0	Economía	NaN	NaN	
8	E5194	NaN	02	0.0	0.0	202266.0	Facultad Economía	D210	3.0	Economía	NaN	NaN	
9	E5194	NaN	02	0.0	0.0	202266.0	Facultad Economía	D553	4.0	Economía	NaN	NaN	

Ilustración 526

Evaluación específica de la unión completa

```
print(f"La unión posee {periodo01.shape[0]} registros y {periodo01.shape[1]} variables.")
La unión posee 41936 registros y 12 variables.

print(f"La unión inicial poseía {union_temp.shape[0]} registros y {union_temp.shape[1]} variables.")
La unión inicial poseía 36361 registros y 10 variables.

print(f"El reporte de calificaciones por carrera posee {perdidas[3].shape[0]} registros y {perdidas[3].shape[1]} variables.")
print(f"El reporte de calificaciones por carrera posee {calificaciones[3].shape[0]} registros y {calificaciones[3].shape[1]} variables.")
print(f"El reporte de becas posee {becas[3].shape[0]} registros y {becas[3].shape[1]} variables.")

El reporte de calificaciones por carrera posee 2640 registros y 8 variables.
El reporte de calificaciones por carrera posee 35056 registros y 6 variables.
El reporte de becas posee 8255 registros y 5 variables.
```

De manera general, se puede concluir que a pesar de que se presenten algunos registros adicionales a los aproximados teóricamente para la unión, el resultado de esta se encuentra dentro de un rango real y posible.

En resumen, se ha logrado realizar el primer proceso de unificación, el de reportes dado el periodo, sus resultados pueden visualizarse a continuación:

Ilustración 527

Vista del primero periodo unificado

Periodo 2021-01

periodo01.head()												
	estudiante	etnia	nivel	segundas_matriculas	terceras_matriculas	periodo	facultad	docente	area_asignatura	carrera	aprobado	creditos
0	E9532	NaN	01	1.0	0.0	202161.0	Dirección Pastoral Univer	D776	2.0	Enfermería	NaN	NaN
1	E9532	NaN	01	1.0	0.0	202161.0	Facultad Cs Exactas Naturales	D699	5.0	Enfermería	NaN	NaN
2	E9532	NaN	01	1.0	0.0	202161.0	Facultad Enfermería	D141	9.0	Enfermería	NaN	NaN
3	E9532	NaN	01	1.0	0.0	202161.0	Facultad Enfermería	D152	3.0	Enfermería	NaN	NaN
4	E2371	Mestizo	07	0.0	0.0	202161.0	Facultad Economía	D177	3.0	Economía	NaN	NaN

Ilustración 528

Vista del segundo periodo unificado

Periodo 2021-02

periodo02.head()												
	estudiante	etnia	nivel	segundas_matriculas	terceras_matriculas	periodo	facultad	docente	area_asignatura	carrera	aprobado	creditos
0	E2371	Mestizo	08	0.0	0.0	202166.0	Facultad Economía	D579	3.0	Economía	NaN	NaN
1	E2371	Mestizo	08	0.0	0.0	202166.0	Facultad Economía	D126	3.0	Economía	NaN	NaN
2	E2371	Mestizo	08	0.0	0.0	202166.0	Facultad Economía	D682	3.0	Economía	NaN	NaN
3	E2371	Mestizo	08	0.0	0.0	202166.0	Facultad Economía	D639	3.0	Economía	NaN	NaN
4	E2371	Mestizo	08	0.0	0.0	202166.0	Facultad Economía	D728	3.0	Economía	NaN	NaN

Ilustración 529

Vista del tercer periodo unificado

Periodo 2022-01

```
periodo03.head()
```

	estudiante	etnia	nivel	segundas_matriculas	terceras_matriculas	periodo	facultad	docente	area_asignatura	carrera	aprobado	creditos
0	E2838	NaN	06	0.0	0.0	202261.0	Facultad Cs Adm. y Contables	D199	4.0	Negocios Internacionales	NaN	NaN
1	E2838	NaN	06	0.0	0.0	202261.0	Facultad Cs Adm. y Contables	D267	4.0	Negocios Internacionales	NaN	NaN
2	E2838	NaN	06	0.0	0.0	202261.0	Facultad Cs Adm. y Contables	D213	4.0	Negocios Internacionales	NaN	NaN
3	E2838	NaN	06	0.0	0.0	202261.0	Facultad Cs Adm. y Contables	D758	4.0	Negocios Internacionales	NaN	NaN
4	E2838	NaN	06	0.0	0.0	202261.0	Facultad Cs Adm. y Contables	D439	4.0	Negocios Internacionales	NaN	NaN

Ilustración 530

Vista del cuarto periodo unificado

Periodo 2022-02

```
periodo04.head()
```

	estudiante	etnia	nivel	segundas_matriculas	terceras_matriculas	periodo	facultad	docente	area_asignatura	carrera	aprobado	creditos
0	E2838	NaN	07	0.0	0.0	202266.0	Facultad Cs Adm. y Contables	D884	4.0	Negocios Internacionales	N	0.0
1	E2838	NaN	07	0.0	0.0	202266.0	Facultad Cs Adm. y Contables	D43	4.0	Negocios Internacionales	N	0.0
2	E2838	NaN	07	0.0	0.0	202266.0	Facultad Cs Adm. y Contables	D73	4.0	Negocios Internacionales	N	0.0
3	E2838	NaN	07	0.0	0.0	202266.0	Facultad Cs Adm. y Contables	D153	4.0	Negocios Internacionales	N	0.0
4	E2838	NaN	07	0.0	0.0	202266.0	Facultad Cs Adm. y Contables	D104	4.0	Negocios Internacionales	N	0.0

Preprocesamiento de los conjuntos de datos de los periodos unificados

Como se ha mencionado, antes de unir los conjuntos de datos de los diferentes periodos académicos, es crucial realizar un segundo preprocesamiento sobre las uniones previamente realizadas, para empezar, se identificarán las variables existentes en los periodos unificados:

Ilustración 531

Código que permite la identificación de variables

```
# Impresión de las variables del periodo 2021-01
print("Variables en el periodo 2021-01:")
for variable in periodo01.columns:
    print(f"\t-{variable}")

# Impresión de las variables del periodo 2021-02
print("\nVariables en el periodo 2021-02:")
for variable in periodo02.columns:
    print(f"\t-{variable}")

# Impresión de las variables del periodo 2022-01
print("\nVariables en el periodo 2022-01:")
for variable in periodo03.columns:
    print(f"\t-{variable}")

# Impresión de las variables del periodo 2022-02
print("\nVariables en el periodo 2022-02:")
for variable in periodo04.columns:
    print(f"\t-{variable}")
```

Ilustración 532

Resultados del primer y segundo periodos unificados

Variables en el periodo 2021-01:

- estudiante
- etnia
- nivel
- segundas_matriculas
- terceras_matriculas
- periodo
- facultad
- docente
- area_asignatura
- carrera
- aprobado
- creditos

Variables en el periodo 2021-02:

- estudiante
- etnia
- nivel
- segundas_matriculas
- terceras_matriculas
- periodo
- facultad
- docente
- area_asignatura
- carrera
- aprobado
- creditos

Ilustración 533

Resultados del tercer y cuarto periodos unificados

Variables en el periodo 2022-01:

- estudiante
- etnia
- nivel
- segundas_matriculas
- terceras_matriculas
- periodo
- facultad
- docente
- area_asignatura
- carrera
- aprobado
- creditos

Variables en el periodo 2022-02:

- estudiante
- etnia
- nivel
- segundas_matriculas
- terceras_matriculas
- periodo
- facultad
- docente
- area_asignatura
- carrera
- aprobado
- creditos

A partir de la identificación de todas las variables, es necesario realizar un análisis de compatibilidad entre periodos en términos de tipos de datos, por lo que se procederá a realizar una inspección visual de la siguientes comparativas:

Ilustración 534

Comparativa del primer periodo unificado

```
# Periodo 2021-01
print("Variables del periodo 2021-01:")
periodo01.dtypes
```

```
Variables del periodo 2021-01:
estudiante          object
etnia               object
nivel              object
segundas_matriculas float64
terceras_matriculas float64
periodo            float64
facultad           object
docente            object
area_asignatura     float64
carrera            object
aprobado           object
creditos           float64
dtype: object
```

Ilustración 535

Comparativa del segundo periodo unificado

```
: # Periodo 2021-02
print("Variables del periodo 2021-02:")
periodo02.dtypes
```

```
Variables del periodo 2021-02:
: estudiante          object
  etnia               object
  nivel              object
  segundas_matriculas float64
  terceras_matriculas float64
  periodo            float64
  facultad           object
  docente            object
  area_asignatura     float64
  carrera            object
  aprobado           object
  creditos           float64
dtype: object
```

Ilustración 536

Comparativa del tercer periodo unificado

```
: # Periodo 2022-01
print("Variables del periodo 2022-01:")
periodo03.dtypes

Variables del periodo 2022-01:
: estudiante          object
  etnia              object
  nivel             object
  segundas_matriculas float64
  terceras_matriculas float64
  periodo           float64
  facultad          object
  docente           object
  area_asignatura    float64
  carrera           object
  aprobado          object
  creditos          float64
dtype: object
```

Ilustración 537

Comparativa del primer periodo unificado

```
# Periodo 2022-02
print("Variables del periodo 2022-02:")
periodo04.dtypes

Variables del periodo 2022-02:
estudiante          object
etnia              object
nivel             object
segundas_matriculas float64
terceras_matriculas float64
periodo           float64
facultad          object
docente           object
area_asignatura    float64
carrera           object
aprobado          object
creditos          float64
dtype: object
```

En resumen, se puede apreciar las siguientes conclusiones de la comparativa realizada:

Tabla 38

Detalle de variables a partir de la comparativa

Variables	¿Requiere armonización?
estudiante	No
etnia	No
nivel	No
segundas_matriculas	No
terceras_matriculas	No
periodo	No
facultad	No
docente	No
area_asignatura	No
carrera	No
aprobado	No
creditos	No

A partir de la visualización de la tabla, se puede concluir que los conjuntos de datos de los periodos unificados no requieren de ningún preprocesamiento y se encuentran listos para su unificación.

Unificación de los conjuntos de datos

En esta sección se llevará a cabo el proceso de unificación de los conjuntos de datos, sin embargo, hay que notar que se realizará la unión por concatenación, es decir, no se entrelazarán los conjuntos de datos, si no, no palabras simples, se colocará uno después del otro. El proceso de unificación consta de un solo código con sus evaluaciones, los siguientes serán visibles a continuación:

Ilustración 538

Concatenación de todos los periodos

```
# Se enlistan los periodos existentes
periodos = [periodo01, periodo02, periodo03, periodo04]

# Se realiza la concatenación
datos = pd.concat(periodos, axis=0)
```

Ilustración 539

Evaluación visual de la unión

datos.head()												
	estudiante	etnia	nivel	segundas_matriculas	terceras_matriculas	periodo	facultad	docente	area_asignatura	carrera	aprobado	creditos
0	E9532	NaN	01	1.0	0.0	202161.0	Dirección Pastoral Univer	D776	2.0	Enfermería	NaN	NaN
1	E9532	NaN	01	1.0	0.0	202161.0	Facultad Cs Exactas Naturales	D699	5.0	Enfermería	NaN	NaN
2	E9532	NaN	01	1.0	0.0	202161.0	Facultad Enfermería	D141	9.0	Enfermería	NaN	NaN
3	E9532	NaN	01	1.0	0.0	202161.0	Facultad Enfermería	D152	3.0	Enfermería	NaN	NaN
4	E2371	Mestizo	07	0.0	0.0	202161.0	Facultad Economía	D177	3.0	Economía	NaN	NaN

Ilustración 540

Evaluación avanzada de la unión

```
print(f"La unión posee {datos.shape[0]} registros y {datos.shape[1]} variables.")
La unión posee 162152 registros y 12 variables.

print(f"El periodo 2021-01 poseía {periodo01.shape[0]} registros y {periodo01.shape[1]} variables.")
print(f"El periodo 2021-02 poseía {periodo02.shape[0]} registros y {periodo02.shape[1]} variables.")
print(f"El periodo 2022-01 poseía {periodo03.shape[0]} registros y {periodo03.shape[1]} variables.")
print(f"El periodo 2022-02 poseía {periodo04.shape[0]} registros y {periodo04.shape[1]} variables.")

El periodo 2021-01 poseía 41936 registros y 12 variables.
El periodo 2021-02 poseía 41253 registros y 12 variables.
El periodo 2022-01 poseía 40253 registros y 12 variables.
El periodo 2022-02 poseía 38710 registros y 12 variables.

print("El total de todos los periodos debería ser de:", periodo01.shape[0] + periodo02.shape[0] + periodo03.shape[0] + periodo04.shape[0] )
El total de todos los periodos debería ser de: 162152
```

En síntesis, a partir de la evaluación es visible que la unión de todos los periodos corresponde al total de todos los tamaños de cada periodo, esto indica que la unión fue un éxito. Por lo que, después de seguir cuidadosamente la estrategia de integración de los conjuntos de datos, se puede afirmar con confianza que la unión fue un éxito.

El resultado de la unificación es un conjunto de datos que proporciona una vista completa y detallada de los datos académicos de la Pontificia Universidad Católica del Ecuador. Esta base

de datos consolidada es ahora un recurso valioso que puede utilizarse para el modelo de machine learning planteado en esta investigación.

4.2.6. Vista previa del conjunto de datos unificado

En este corto subcapítulo únicamente se visualizará una vista previa del conjunto de datos unificado:

Ilustración 541

Vista previa del conjunto de datos unificado

datos.head(10)												
	estudiante	etnia	nivel	segundas_matriculas	terceras_matriculas	periodo	facultad	docente	area_asignatura	carrera	aprobado	creditos
0	E9532	NaN	01	1.0	0.0	202161.0	Dirección Pastoral Univer	D776	2.0	Enfermería	NaN	NaN
1	E9532	NaN	01	1.0	0.0	202161.0	Facultad Cs Exactas Naturales	D699	5.0	Enfermería	NaN	NaN
2	E9532	NaN	01	1.0	0.0	202161.0	Facultad Enfermería	D141	9.0	Enfermería	NaN	NaN
3	E9532	NaN	01	1.0	0.0	202161.0	Facultad Enfermería	D152	3.0	Enfermería	NaN	NaN
4	E2371	Mestizo	07	0.0	0.0	202161.0	Facultad Economía	D177	3.0	Economía	NaN	NaN
5	E2371	Mestizo	07	0.0	0.0	202161.0	Facultad Economía	D172	3.0	Economía	NaN	NaN
6	E2371	Mestizo	07	0.0	0.0	202161.0	Facultad Economía	D158	3.0	Economía	NaN	NaN
7	E2371	Mestizo	07	0.0	0.0	202161.0	Facultad Economía	D579	3.0	Economía	NaN	NaN
8	E2371	Mestizo	07	0.0	0.0	202161.0	Facultad Economía	D507	3.0	Economía	NaN	NaN
9	E2371	Mestizo	07	0.0	0.0	202161.0	Facultad Economía	D955	3.0	Economía	NaN	NaN

4.2.7. Limpieza de datos del conjunto de datos unificado

En resumen, hasta este momento se ha llevado a cabo la unión de los diferentes conjuntos de datos y se ha obtenido un conjunto de datos unificado con todos los datos de todos los reportes y las variables más influyentes para el rendimiento académico, ahora, es el momento de limpiar estos datos a través de rigurosos procesos. (Amazon, s.f.) Por lo general, entre las acciones que se llevan a cabo en el proceso de limpieza son: eliminación de duplicados, tratamiento de valores faltantes, normalización y estandarización. (Amazon, s.f.).

Eliminación de duplicados del conjunto de datos unificado

En primer lugar, se llevará a cabo un segundo proceso de eliminación de duplicados debido a que, dado el proceso de unificación, es posible que se hayan registrado valores duplicados, sin embargo, el verdadero motivo de realizar una nueva eliminación es con la finalidad de evitar posibles problemas futuros. El realizo lo siguiente:

Ilustración 542

Eliminación de duplicados del conjunto de datos unificado

```
# Visualización de las dimensiones antes de la eliminación de duplicados
print(f"Existen {datos.shape[0]} registros y {datos.shape[1]} variables en el conjunto de datos unificado.")
Existen 162152 registros y 12 variables en el conjunto de datos unificado.

# Se eliminan los duplicados del periodo
datos.drop_duplicates(inplace=True)

# Visualización de las dimensiones después de la eliminación de duplicados
print(f"Existen {datos.shape[0]} registros y {datos.shape[1]} variables en el conjunto de datos unificado.")
Existen 162007 registros y 12 variables en el conjunto de datos unificado.
```

A partir del proceso, sucedió lo siguiente:

Tabla 39

Resultados de la eliminación de duplicados

Cantidad inicial de registros	Cantidad de registros eliminados	Cantidad actual de registros
162152	145	162007

Identificación de datos faltantes del conjunto de datos unificado

Para poder tratar con los valores faltantes, en un comienzo es necesario identificarlos y evaluar que se debe hacer con ellos debido a que distintas técnicas de tratamiento de valores faltantes pueden generar repercusiones importantes en el modelo de predicción del rendimiento académico si no se les trata adecuadamente. Para empezar, se visualizará la cantidad de registros en el conjunto de datos unificado:

Ilustración 543

Cantidad de registros en el conjunto de datos

```
print(f"Existen {datos.shape[0]} registros y {datos.shape[1]} variables en el conjunto de datos unificado.")
```

Existen 162007 registros y 12 variables en el conjunto de datos unificado.

Conociendo el total de los registros, se pueden analizar las estadísticas descriptivas del conjunto de datos:

Ilustración 544

Estadísticas descriptivas de los datos

```
datos.describe(include=object)
```

	estudiante	etnia	nivel	facultad	docente	carrera	aprobado
count	162007	46608	161564	157868	156754	157868	25928
unique	12097	8	16	15	960	80	1
top	E3649	Mestizo	01	Facultad Enfermería	D38	Derecho	N
freq	48	43763	21595	22544	849	23664	25928

```
datos.describe()
```

	segundas_matriculas	terceras_matriculas	periodo	area_asignatura	creditos
count	160893.000000	160893.000000	157868.000000	157868.000000	25928.000000
mean	0.144152	0.013643	202212.001723	5.079395	3.846257
std	0.531342	0.150186	50.020791	2.699660	2.228041
min	-1.000000	-1.000000	202161.000000	1.000000	0.000000
25%	0.000000	0.000000	202161.000000	3.000000	3.000000
50%	0.000000	0.000000	202166.000000	4.000000	4.000000
75%	0.000000	0.000000	202261.000000	7.000000	5.000000
max	6.000000	5.000000	202266.000000	11.000000	20.000000

A partir de las estadísticas descriptivas, se puede notar la existe de algunas variables con valores faltantes, por lo que se procederá a realizar un análisis de la cantidad y proporción de valores faltantes, los códigos y sus resultados se admirar a continuación:

Ilustración 545

Análisis de cantidad de valores faltantes

```
# Se muestra una sumarización de valores faltantes por variable
print("Existen las siguientes cantidades de valores faltantes: ")
datos.isnull().sum()
```

Existen las siguientes cantidades de valores faltantes:

```
estudiante      0
etnia           115399
nivel           443
segundas_matriculas  1114
terceras_matriculas  1114
periodo         4139
facultad        4139
docente         5253
area_asignatura  4139
carrera         4139
aprobado        136079
creditos        136079
dtype: int64
```

Ilustración 546

Análisis del porcentaje de los valores faltantes

```
print("Existen los siguientes porcentajes de valores faltantes:\n")

# Se almacena el total de registros de los datos
total = datos.shape[0]

# Se itera por todas las variables
for variable in datos.columns:
    # Se recopila su total de valores nulos
    faltantes = datos[variable].isnull().sum()

    # Se calcula el porcentaje
    porcentaje = round(faltantes/total*100, 2)

    # Se calcula el equivalente
    equivalente = int(round(total * (porcentaje/100), 0))

    # Se imprime el resultado
    print(f"\nAnalizando la variable '{variable}'...")
    print(f"\tPosee una cantidad de {faltantes} valores faltantes.")
    print(f"\tSu porcentaje de valores faltantes es del: {porcentaje}%")
    print(f"\tCon su eliminación se dispondría de {total - equivalente} registros.")
```

Ilustración 547

Pare 1 de los resultados del análisis del porcentaje de los valores faltantes

```
Analizando la variable 'estudiante'...
  Posee una cantidad de 0 valores faltantes.
  Su porcentaje de valores faltantes es del: 0.0%
  Con su eliminación se dispondría de 162007 registros.

Analizando la variable 'etnia'...
  Posee una cantidad de 115399 valores faltantes.
  Su porcentaje de valores faltantes es del: 71.23%
  Con su eliminación se dispondría de 46609 registros.

Analizando la variable 'nivel'...
  Posee una cantidad de 443 valores faltantes.
  Su porcentaje de valores faltantes es del: 0.27%
  Con su eliminación se dispondría de 161570 registros.

Analizando la variable 'segundas_matriculas'...
  Posee una cantidad de 1114 valores faltantes.
  Su porcentaje de valores faltantes es del: 0.69%
  Con su eliminación se dispondría de 160889 registros.

Analizando la variable 'terceras_matriculas'...
  Posee una cantidad de 1114 valores faltantes.
  Su porcentaje de valores faltantes es del: 0.69%
  Con su eliminación se dispondría de 160889 registros.

Analizando la variable 'periodo'...
  Posee una cantidad de 4139 valores faltantes.
  Su porcentaje de valores faltantes es del: 2.55%
  Con su eliminación se dispondría de 157876 registros.
```

Ilustración 548

Parte final de los resultados del análisis del porcentaje de los valores faltantes

```
Analizando la variable 'facultad'...
  Posee una cantidad de 4139 valores faltantes.
  Su porcentaje de valores faltantes es del: 2.55%
  Con su eliminación se dispondría de 157876 registros.

Analizando la variable 'docente'...
  Posee una cantidad de 5253 valores faltantes.
  Su porcentaje de valores faltantes es del: 3.24%
  Con su eliminación se dispondría de 156758 registros.

Analizando la variable 'area_asignatura'...
  Posee una cantidad de 4139 valores faltantes.
  Su porcentaje de valores faltantes es del: 2.55%
  Con su eliminación se dispondría de 157876 registros.

Analizando la variable 'carrera'...
  Posee una cantidad de 4139 valores faltantes.
  Su porcentaje de valores faltantes es del: 2.55%
  Con su eliminación se dispondría de 157876 registros.

Analizando la variable 'aprobado'...
  Posee una cantidad de 136079 valores faltantes.
  Su porcentaje de valores faltantes es del: 84.0%
  Con su eliminación se dispondría de 25921 registros.

Analizando la variable 'creditos'...
  Posee una cantidad de 136079 valores faltantes.
  Su porcentaje de valores faltantes es del: 84.0%
  Con su eliminación se dispondría de 25921 registros.
```

En síntesis, tras el análisis realizado se pudo determinar lo siguiente:

Tabla 40

Identificación de valores faltantes

Variable	Cantidad de valores faltantes	Porcentaje de valores faltantes	Cantidad total de registros en caso de eliminación de valores faltantes
estudiante	0	0%	162007
etnia	115399	71.23%	46609
nivel	443	0.27%	161570
segundas_matriculas	1114	0.69%	160889
terceras_matriculas	1114	0.69%	160889

periodo	4139	2.55%	157876
facultad	4139	2.55%	157876
docente	5253	3.24%	156758
area_asignatura	4139	2.55%	157876
carrera	4139	2.55%	157876
aprobado	136079	84.0%	25921
creditos	136079	84.0%	25921

En este sentido, una vez recapitulado la información de los valores faltantes de manera general, es momento de realizar un análisis específico de las variables.

Análisis de la variable estudiante. la variable al tener 0 valores faltantes no se debe realizar ninguna acción.

Análisis de la variable etnia. La variable "etnia" presenta un porcentaje significativo de valores faltantes, exactamente un 71.23%, este alto porcentaje indica que más de dos tercios de los datos de esta variable no están disponibles en el conjunto de datos unificado.

Un valor tan elevado de datos faltantes puede ser motivo de preocupación ya que esto puede afectar la calidad y la precisión del análisis, sobre todo, porque, dado que la variable "etnia" podría ser relevante para el estudio del rendimiento académico, la falta de datos en esta área puede limitar la comprensión de cómo este factor afecta el rendimiento de los estudiantes.

La alta tasa de valores faltantes plantea la cuestión de cómo se debe manejar esta variable en el análisis, ya que, dependiendo de las circunstancias y de los objetivos del análisis, puede ser apropiado imputar los valores faltantes, eliminar esta variable del análisis o realizar un análisis de sensibilidad para comprender cómo la falta de datos en esta variable puede afectar las conclusiones del estudio.

No obstante, al observar el impacto que tendría la eliminación de todos los valores faltantes de la variable "etnia", se puede notar que el conjunto de datos aún contaría con un total de 46609 registros. Con base en esta situación, realmente existen dos alternativas viables a considerar:

1. Desarrollar el modelo de aprendizaje automático sin incluir la variable "etnia". Esto implicaría descartar esta variable de los datos debido a su alto porcentaje de datos faltantes, y enfocarse en las otras variables disponibles que estén más completas y que puedan aportar a la predicción del rendimiento académico.
2. Imputar valores en los valores faltantes de la variable "etnia". Esto implicaría evaluar en qué medida se escogerían los valores para rellenar los valores faltantes.
3. Eliminar los valores faltantes de la variable "etnia" y desarrollar el modelo de aprendizaje automático con el conjunto resultante de 46609 registros. Esta opción podría conllevar el riesgo de sesgos o imprecisiones debido a la cantidad de información de la variable "etnia" que se ha perdido.

A partir de lo considerado, se ha tomado la decisión de eliminar los valores faltantes de la variable "etnia" debido a las siguientes razones:

- La variable "etnia" tiene una influencia significativa en el rendimiento académico a cierto nivel. Su inclusión puede mejorar la precisión del modelo al capturar patrones y correlaciones no evidentes en otras variables.
- Sin embargo, con un 71.23% de los datos faltantes para esta variable, la imputación podría introducir un sesgo considerable y es probable que no sea la opción más adecuada.

- Este es un modelo con 12 variables, y un conjunto de datos de 46,609 observaciones es razonablemente grande. Es muy probable que sea suficiente para desarrollar un modelo de predicción del rendimiento académico efectivo.

Análisis de la variable nivel. Dado que los datos faltantes de esta variable representan únicamente el 0.27% de los registros, se ha decidido optar por una imputación de valores dada la mediana o el valor más frecuente de los datos de esta variable.

Análisis de las variables segundas_matriculas y terceras_matriculas. Se puede observar que existen en ambos casos un 0.69% de valores faltantes, dado el mínimo porcentaje en comparación a la enorme cantidad de registros, se ha decidido realizar una imputación con respecto a la mediana de ambas variables.

Análisis de las variables periodo, facultad, area_asignatura y carrera. En este caso, es posible observar que en todas las variables hay 4139 valores faltantes, lo cual equivale al 2.55% de los 162007 datos totales en el conjunto de datos unificado, estos valores faltantes corresponden específicamente a los datos que estaban presentes en un reporte, pero no en los demás. Para aclarar esta situación, se realizará un análisis del contenido de estas variables a continuación.

Ilustración 549

Análisis de las variables periodo, facultad, area_asignatura y carrera

```
# tomamos los valores iguales entre los valores nulos de las diferentes variables
evaluacion = datos["estudiante"][datos["facultad"].isnull()] == datos["estudiante"][datos["periodo"].isnull()]
evaluacion = evaluacion == datos["estudiante"][datos["area_asignatura"].isnull()]
evaluacion = evaluacion == datos["estudiante"][datos["carrera"].isnull()]

# Evaluamos si el total de los registros faltantes iguales a través de todos los periodos corresponden a 4139
print("Todos los registros nulos en las diferentes variables son los mismos?", len(evaluacion) == 4139)
```

Como se puede observar, los registros faltantes en cada una de estas variables corresponden a los mismos registros, además, dado que estos representan sólo el 2.55% del total

de los datos, se ha tomado la decisión de eliminar estos registros, porque esta acción ayudará a evitar la introducción de sesgos que podrían resultar de la imputación de estos valores faltantes y debería tener un impacto mínimo en la capacidad del modelo resultante para hacer predicciones precisas y útiles.

Análisis de la variable docente. Se puede observar que hay un 3.24% de valores faltantes en esta variable, que se refiere a docentes específicos, además, por lo que la imputación de estos valores podría introducir sesgos irreales, se ha tomado la decisión de eliminar estos registros. Este enfoque es prudente, ya que intentar imputar datos sobre docentes específicos podría distorsionar los resultados y afectar la precisión del modelo de predicción.

Análisis de la variable aprobado. En esta ocasión, se observa un total del 84% de valores faltantes, lo que podría parecer alarmante a primera vista, sin embargo, es importante entender que estos datos representan a estudiantes que han aprobado sus asignaturas, en otras palabras, la ausencia de información en estos casos se interpreta como una indicación de aprobación. Por lo tanto, se tomará la decisión de imputar el valor "S" en todos los registros que no tienen un valor en esta variable, reflejando así el estado de aprobación de los estudiantes.

Análisis de la variable créditos. En esta ocasión, también se observa un alto porcentaje de valores faltantes, sin embargo, esto es esperado, ya que esta variable sólo está presente en el reporte de estudiantes reprobados; así que, dado que se dispone de 25921 registros que contienen información sobre los créditos de asignaturas en todas las áreas (como fue demostrado en la fase anterior de comprensión de los datos), se ha decidido imputar a la mediana de los créditos que contiene cada área de asignatura. Este enfoque permite utilizar la información disponible para llenar los espacios vacíos de una manera que es coherente con los patrones observados en los datos existentes.

Tratamiento de datos faltantes

Una vez identificados los datos faltantes se puede proceder a tratarlos, para esto, dependiendo de la naturaleza y el volumen de los datos faltantes, se puede optar por varias estrategias. La primera estrategia es la imputación de datos, que consiste en reemplazar los valores faltantes por otros valores estimados, las variables que utilizarán esta estrategia son:

- nivel
- segundas_matriculas
- terceras_matriculas
- aprobado
- creditos

La segunda estrategia es la eliminación de registros con datos faltantes, esta puede ser una opción viable si la cantidad de datos faltantes es pequeña y la eliminación de estos registros no tendría un impacto significativo en la calidad del análisis, sin embargo, existen factores adicionales como los descritos en la anterior sección, las variables que utilizarán esta estrategia son:

- etnia
- periodo
- facultad
- area_asignatura
- carrera
- docente

Por otro lado, en esta sección se creará un nuevo conjunto de datos denominado `datos_procesados`, en él se llevará a cabo todo el proceso de limpieza de datos acordado, con el objetivo de preservar la versión original del conjunto de datos para que los investigadores que tengan acceso a este trabajo de titulación puedan utilizar estos datos y considerar diferentes decisiones a las que se han tomado en esta investigación. En otras palabras, usted estimado lector podrá tener acceso a los datos sin procesar para que si usted desea tome decisiones diferentes que las realizadas en este trabajo de titulación.

A continuación, se podrán visualizar las capturas respectivas a cada proceso llevado a cabo para el tratamiento de valores faltantes:

Ilustración 550

Creación de un nuevo conjunto de datos y eliminación de valores faltantes de la variable *etnia*

6.3.1. Creación del nuevo conjunto de datos

```
# Se crea el nuevo conjunto de datos mencionado al comienzo de la sección
datos_procesados = datos.copy()
```

6.3.2. Eliminación de valores faltantes

Eliminación de valores faltantes de la variable *etnia*

```
# Se muestra las dimensiones de los datos antes de la eliminación
print(f"Existen {datos_procesados.shape[0]} registros y {datos_procesados.shape[1]} variables en el conjunto de datos unificado.")
```

Existen 162007 registros y 12 variables en el conjunto de datos unificado.

```
# Se eliminan los valores faltantes de la variable etnia ~ 71.23% de los datos
datos_procesados = datos_procesados.dropna(subset="etnia")
print("Se ha realizado la eliminación con éxito!")
```

Se ha realizado la eliminación con éxito!

```
# Se muestra las dimensiones de los datos después de la eliminación ~ 115399 registros eliminados en total
print(f"Existen {datos_procesados.shape[0]} registros y {datos_procesados.shape[1]} variables en el conjunto de datos unificado.")
```

Existen 46608 registros y 12 variables en el conjunto de datos unificado.

Ilustración 551

Eliminación de valores faltantes de las variables periodo, facultad, area_asignatura y carrera

```
# Se muestra las dimensiones de los datos antes de la eliminación
print(f"Existen {datos_procesados.shape[0]} registros y {datos_procesados.shape[1]} variables en el conjunto de datos unificado.")

Existen 46608 registros y 12 variables en el conjunto de datos unificado.

# Se eliminan los valores faltantes de las variables mencionadas ~ 2.55% de los datos
datos_procesados = datos_procesados.dropna(subset=["periodo"])
print("Se ha realizado la eliminación con éxito!")

Se ha realizado la eliminación con éxito!

# Se muestra las dimensiones de los datos después de la eliminación ~ 4139 registros eliminados en total
print(f"Existen {datos_procesados.shape[0]} registros y {datos_procesados.shape[1]} variables en el conjunto de datos unificado.")

Existen 45973 registros y 12 variables en el conjunto de datos unificado.
```

Ilustración 552

Eliminación de valores faltantes de la variable docente

Eliminación de valores faltantes de la variable docente

```
# Se muestra las dimensiones de los datos antes de la eliminación
print(f"Existen {datos_procesados.shape[0]} registros y {datos_procesados.shape[1]} variables en el conjunto de datos unificado.")

Existen 45973 registros y 12 variables en el conjunto de datos unificado.

# Se eliminan los valores faltantes de la variable etnia ~ 3.24% de los datos
datos_procesados = datos_procesados.dropna(subset=["docente"])
print("Se ha realizado la eliminación con éxito!")

Se ha realizado la eliminación con éxito!

# Se muestra las dimensiones de los datos después de la eliminación ~ 45973 registros en total
print(f"Existen {datos_procesados.shape[0]} registros y {datos_procesados.shape[1]} variables en el conjunto de datos unificado.")

Existen 45973 registros y 12 variables en el conjunto de datos unificado.
```

Ilustración 553

Verificación del proceso de eliminación de valores faltantes

```
# Se muestra una sumaria de valores faltantes por variable
print("Existen las siguientes cantidades de valores faltantes: ")
datos_procesados.isnull().sum()

Existen las siguientes cantidades de valores faltantes:
estudiante          0
etnia                0
nivel               110
segundas_matriculas 0
terceras_matriculas 0
periodo             0
facultad            0
docente             0
area_asignatura     0
carrera             0
aprobado            38377
creditos            38377
dtype: int64
```

Es importante notar que se puede confirmar que las variables consideradas para la eliminación de valores faltantes fueron realizadas con éxito, además, es visible que ya no

es necesario tratar los valores faltantes de las variables segundas_matriculas y terceras_matriculas. A continuación, se proseguirá con la técnica de imputación de datos.

Ilustración 554

Parte 1 de la imputación de valores faltantes en la variable nivel

Imputación de valores faltantes en la variable nivel

```
# Se muestra una sumarización de valores faltantes por variable antes del cambio
print("Existen las siguientes cantidades de valores faltantes: ")
datos_procesados.isnull().sum()
```

```
Existen las siguientes cantidades de valores faltantes:
estudiante          0
etnia                0
nivel               110
segundas_matriculas 0
terceras_matriculas 0
periodo             0
facultad            0
docente             0
area_asignatura     0
carrera             0
aprobado            38377
creditos            38377
dtype: int64
```

```
# Se muestra el valor más frecuente de la variable nivel
datos_procesados["nivel"].describe()
```

```
count    45863
unique     12
top        08
freq     13830
Name: nivel, dtype: object
```


Ilustración 555

Parte 2 de la imputación de valores faltantes en la variable nivel

```
# Se imputa el nivel 08 en los valores faltantes
datos_procesados["nivel"] = datos_procesados["nivel"].fillna("08")
print("Imputación de valores realizada con éxito!")
```

Imputación de valores realizada con éxito!

```
# Se muestra una sumarización de valores faltantes por variable
print("Existen las siguientes cantidades de valores faltantes: ")
datos_procesados.isnull().sum()
```

Existen las siguientes cantidades de valores faltantes:

estudiante	0
etnia	0
nivel	0
segundas_matriculas	0
terceras_matriculas	0
periodo	0
facultad	0
docente	0
area_asignatura	0
carrera	0
aprobado	38377
creditos	38377

dtype: int64

Ilustración 556

Parte 1 de la imputación de valores faltantes en la variable aprobado

```
# Se muestra una sumarización de valores faltantes por variable antes del cambio
print("Existen las siguientes cantidades de valores faltantes: ")
datos_procesados.isnull().sum()
```

Existen las siguientes cantidades de valores faltantes:

estudiante	0
etnia	0
nivel	0
segundas_matriculas	0
terceras_matriculas	0
periodo	0
facultad	0
docente	0
area_asignatura	0
carrera	0
aprobado	38377
creditos	38377

dtype: int64

```
# Se imputa "S" en los valores faltantes para marcar que los estudiantes han aprobado
datos_procesados["aprobado"] = datos_procesados["aprobado"].fillna("S")
print("Imputación de valores realizada con éxito!")
```

Imputación de valores realizada con éxito!

Ilustración 557

Parte final de la imputación de valores faltantes en la variable aprobado

```
# Se muestra una sumarización de valores faltantes por variable después del cambio
print("Existen las siguientes cantidades de valores faltantes: ")
datos_procesados.isnull().sum()
```

Existen las siguientes cantidades de valores faltantes:

estudiante	0
etnia	0
nivel	0
segundas_matriculas	0
terceras_matriculas	0
periodo	0
facultad	0
docente	0
area_asignatura	0
carrera	0
aprobado	0
creditos	38377

dtype: int64

Ilustración 558

Parte 1 de la imputación de valores faltantes en la variable creditos

```
# Almacenará todas las medianas
medianas = {}

# Se toman todas las áreas existentes
areas = datos_procesados["area_asignatura"].unique()

# Se itera por todas las áreas
for area in areas:
    # Se crea un dataframe filtrado
    temp_df = datos_procesados["creditos"][datos_procesados["area_asignatura"] == area]

    # Se toma la mediana del dataframe filtrado y se guarda en el diccionario medianas
    medianas[int(area)] = temp_df.median()

# Se muestran los resultados
print("Existen las siguientes medianas:")
for area, credito in medianas.items():
    print(f"El área {area} tiene como mediana: {credito} creditos.")
```

Existen las siguientes medianas:
El área 03 tiene como mediana: 3.0 creditos.
El área 07 tiene como mediana: 4.5 creditos.
El área 02 tiene como mediana: 3.0 creditos.
El área 06 tiene como mediana: 4.0 creditos.
El área 09 tiene como mediana: 4.0 creditos.
El área 011 tiene como mediana: 0.0 creditos.
El área 05 tiene como mediana: 3.0 creditos.
El área 04 tiene como mediana: 3.0 creditos.
El área 08 tiene como mediana: nan creditos.
El área 01 tiene como mediana: 2.0 creditos.
El área 010 tiene como mediana: 6.0 creditos.

Ilustración 559

Parte final de la imputación de valores faltantes en la variable creditos

```
# Se imputan las medianas en los valores faltantes
datos_procesados["creditos"] = datos_procesados["creditos"].fillna(datos_procesados["area_asignatura"].map(medianas))
print("Imputación de valores realizada con éxito!")

Imputación de valores realizada con éxito!
```

```
# Se muestra una summarización de valores faltantes por variable
print("Existen las siguientes cantidades de valores faltantes: ")
datos_procesados.isnull().sum()
```

Existen las siguientes cantidades de valores faltantes:

estudiante	0
etnia	0
nivel	0
segundas_matriculas	0
terceras_matriculas	0
periodo	0
facultad	0
docente	0
area_asignatura	0
carrera	0
aprobado	0
creditos	66

dtype: int64

Ilustración 560

Verificación de la imputación de valores faltantes

```
# Se muestra una sumarización de valores faltantes por variable
print("Existen las siguientes cantidades de valores faltantes: ")
datos_procesados.isnull().sum()

Existen las siguientes cantidades de valores faltantes:
estudiante      0
etnia            0
nivel           0
segundas_matriculas  0
terceras_matriculas  0
periodo         0
facultad        0
docente         0
area_asignatura  0
carrera         0
aprobado        0
creditos        66
dtype: int64

cantidad = datos_procesados["creditos"].isnull().sum()
print(f"El porcentaje de valores faltantes de la variable nivel es de: {round(cantidad / datos_procesados.shape[0] * 100, 2)}%")

El porcentaje de valores faltantes de la variable nivel es de: 0.14%
```

En síntesis, a pesar de haberse realizado la imputación, se encontró que persisten 66 valores nulos, sin embargo, esto se debe principalmente a que la mediana identificada en las áreas de asignatura con código "08" resulta también ser un valor faltante; tras esta aclaración, y considerando que la cantidad de valores faltantes es de 66 en un total de 45973 registros, es decir, solo el 0.14%, es factible la idea de proceder con la eliminación de estos valores faltantes.

Ilustración 561

Eliminación de los valores faltantes de la variable creditos

Eliminación de los valores faltantes de la variable creditos

```
# Se muestra las dimensiones de los datos antes de la eliminación
print(f"Existen {datos_procesados.shape[0]} registros y {datos_procesados.shape[1]} variables en el conjunto de datos unificado.")

Existen 45973 registros y 12 variables en el conjunto de datos unificado.

# Se eliminan los valores faltantes de la variable etnia ~ 3.24% de los datos
datos_procesados = datos_procesados.dropna(subset="creditos")
print("Se ha realizado la eliminación con éxito!")

Se ha realizado la eliminación con éxito!

# Se muestra las dimensiones de los datos después de la eliminación ~ 45973 registros en total
print(f"Existen {datos_procesados.shape[0]} registros y {datos_procesados.shape[1]} variables en el conjunto de datos unificado.")

Existen 45907 registros y 12 variables en el conjunto de datos unificado.
```

Ilustración 562

Parte 1 de la revisión final del conjunto de datos

```
# Se muestra las dimensiones de los datos
print(f"Existen {datos_procesados.shape[0]} registros y {datos_procesados.shape[1]} variables en el conjunto de datos unificado.")
```

Existen 45907 registros y 12 variables en el conjunto de datos unificado.

```
# Se muestra una sumariación de valores faltantes por variable
print("Existen las siguientes cantidades de valores faltantes: ")
datos_procesados.isnull().sum()
```

Existen las siguientes cantidades de valores faltantes:

```
estudiante      0
etnia            0
nivel           0
segundas_matriculas  0
terceras_matriculas  0
periodo         0
facultad        0
docente         0
area_asignatura  0
carrera         0
aprobado        0
creditos        0
dtype: int64
```

Ilustración 563

Parte final de la revisión final del conjunto de datos

```
# Se muestra una vista previa
datos_procesados.head()
```

	estudiante	etnia	nivel	segundas_matriculas	terceras_matriculas	periodo	facultad	docente	area_asignatura	carrera	aprobado	creditos
4	E2371	Mestizo	07	0.0	0.0	202161.0	Facultad Economía	D177	3.0	Economía	S	3.0
5	E2371	Mestizo	07	0.0	0.0	202161.0	Facultad Economía	D172	3.0	Economía	S	3.0
6	E2371	Mestizo	07	0.0	0.0	202161.0	Facultad Economía	D158	3.0	Economía	S	3.0
7	E2371	Mestizo	07	0.0	0.0	202161.0	Facultad Economía	D579	3.0	Economía	S	3.0
8	E2371	Mestizo	07	0.0	0.0	202161.0	Facultad Economía	D507	3.0	Economía	S	3.0

En este punto, se puede concluir que, a pesar de la gran cantidad de registros eliminados, el conjunto de datos aún tiene una cantidad suficiente de datos para generar un modelo de predicción de rendimiento académico adecuado, por otra parte, con respecto al proceso de tratamiento de valores faltantes, se puede determinar que ya no existen valores faltantes.

Corrección de inconsistencias de datos

En esta sección se verificarán si existen inconsistencias en los datos del conjunto de datos procesado, a partir de las conclusiones de la verificación se tomarán acciones respectivas. A continuación, se mostrarán los procesos realizados y se comentarán brevemente lo que sucedió en cada ocasión.

Ilustración 564

Verificación y corrección de las variables estudiante y periodo

```
# Eliminación de las variables mencionadas
datos_procesados.drop(columns=["estudiante", "periodo"], axis=0, inplace=True)

# Vista previa de la corrección
datos_procesados.head()
```

	etnia	nivel	segundas_matriculas	terceras_matriculas	facultad	docente	area_asignatura	carrera	aprobado	creditos
4	Mestizo	07	0.0	0.0	Facultad Economía	D177	3.0	Economía	S	3.0
5	Mestizo	07	0.0	0.0	Facultad Economía	D172	3.0	Economía	S	3.0
6	Mestizo	07	0.0	0.0	Facultad Economía	D158	3.0	Economía	S	3.0
7	Mestizo	07	0.0	0.0	Facultad Economía	D579	3.0	Economía	S	3.0
8	Mestizo	07	0.0	0.0	Facultad Economía	D507	3.0	Economía	S	3.0

En específico, con respecto a la variable estudiante, dado que el enfoque del modelo es el determinar si un estudiante va a aprobar o reprobar una asignatura utilizando datos históricos de todos los estudiantes que han cursado por esa misma asignatura, resulta irrelevante conocer el pasado histórico de un determinado estudiante, por lo que se procedió con su eliminación en conjunto con la variable periodo, ya que, el propósito de esta última era únicamente el diferenciar un periodo del otro.

Ilustración 565

Verificación y corrección de la variable etnia

```
# Se visualizan los datos únicos de la variable
print("La variable cuenta con los siguientes datos únicos: ")
for dato in datos_procesados["etnia"].unique():
    print(f"\t-{dato}")
```

```
La variable cuenta con los siguientes datos únicos:
\t-Mestizo
\t-Otro
\t-Blanco
\t-Afroecuatoriano
\t-Mulato
\t-Negro
\t-Indígena
\t-Montubio
```

Sin inconsistencias.

Ilustración 566

Verificación y corrección de la variable nivel

```
# Se visualizan los datos únicos de la variable
print("La variable cuenta con los siguientes datos únicos: ")
for dato in datos_procesados["nivel"].unique():
    print(f"\t-{dato}")
```

```
La variable cuenta con los siguientes datos únicos:
\t-07
\t-10
\t-06
\t-08
\t-05
\t-04
\t-09
\t-03
\t-02
\t-01
\t-EL
\t-ELECT
```

Se puede observar que sí existen inconsistencias en los datos con respecto a "EL" y "ELECT" que muy probablemente tengan un mismo sentido o significado.

Ilustración 567

Análisis avanzado de la variable nivel

Análisis avanzado de la variable

```
cantidad_el = len(datos_procesados[datos_procesados["nivel"] == "EL"])
print(f"Existen {cantidad_el} estudiante con nivel 'EL'")

cantidad_elect = len(datos_procesados[datos_procesados["nivel"] == "ELECT"])
print(f"\nExisten {cantidad_elect} estudiante con nivel 'ELECT'")
```

```
Existen 37 estudiante con nivel 'EL'
```

```
Existen 44 estudiante con nivel 'ELECT'
```

Se identificó la cantidad de registros con dichos valores.

Ilustración 568

Eliminación de los registros con 'EL' y 'ELECT'

Eliminación de los registros con 'EL' y 'ELECT'

```
# Se generan las condiciones que cumplen lo que no deseamos
condicion_el = datos_procesados["nivel"] == "EL"
# Se seleccionan los datos que no cumplen las condiciones
datos_procesados = datos_procesados[~condicion_el]

# Se generan las condiciones que cumplen lo que no deseamos
condicion_elect = datos_procesados["nivel"] == "ELECT"
# Se seleccionan los datos que no cumplen las condiciones
datos_procesados = datos_procesados[~condicion_elect]
```

Dado que existe una inmensa cantidad de registros en comparación a los 37 y 44 estudiantes con registros de nivel irregulares, e, incluyendo que se desconoce el significado de "EL" como de "ELECT", se ha decidido eliminar estos registros.

Ilustración 569

Vista previa del cambio

Vista previa del cambio

```
# Se muestra las dimensiones de los datos
print(f"Existen {datos_procesados.shape[0]} registros y {datos_procesados.shape[1]} variables en el conjunto de datos unificado.")
Existen 45826 registros y 10 variables en el conjunto de datos unificado.

datos_procesados["nivel"].unique()
array(['07', '10', '06', '08', '05', '04', '09', '03', '02', '01'],
      dtype=object)
```

En resumen, se eliminaron los registros con “EL” y “ELECT”, sin embargo, no representa un problema debido a la mínima cantidad de registros que poseían estos valores.

Ilustración 570

Verificación de las variables de segundas_matriculas y terceras_matriculas

```
# Se visualizan los datos únicos de la variable  
print("La variable cuenta con los siguientes datos únicos: ")  
for dato in datos_procesados["segundas_matriculas"].unique():  
    print(f"\t-{dato}")
```

La variable cuenta con los siguientes datos únicos:

-0.0
-1.0
-2.0
-3.0
-5.0
-4.0
-6.0

```
# Se visualizan los datos únicos de la variable  
print("La variable cuenta con los siguientes datos únicos: ")  
for dato in datos_procesados["terceras_matriculas"].unique():  
    print(f"\t-{dato}")
```

La variable cuenta con los siguientes datos únicos:

-0.0
-1.0
-2.0
-3.0
-5.0

Sin inconsistencias.

Ilustración 571

Verificación de la variable facultad

```
# Se visualizan Los datos únicos de La variable
print("La variable cuenta con los siguientes datos únicos: ")
for dato in datos_procesados["facultad"].unique():
    print(f"\t-{dato}")
```

La variable cuenta con los siguientes datos únicos:

- Facultad Economía
- Facultad Arquitectónica Diseño y Artes
- Dirección Pastoral Univer
- Facultad Cs Humanas
- Facultad Ingeniería
- Facultad Enfermería
- Facultad Psicología
- Facultad Com Lingüística y Lit
- Facultad Cs Adm. y Contables
- Facultad Jurisprudencia
- Facultad Medicina
- Facultad Cs Exactas Naturales
- Facultad Cs Filosófico Teológico
- Facultad Cs de la Educación

Sin inconsistencias.

Ilustración 572

Verificación de la variable docente

```
# Se toma una lista de todos los docentes únicos
docentes = datos_procesados["docente"].unique()

# Variable que evaluará si existen inconsistencias
verificadora = True

# Verifica si algún docente posee un nombre equivocado
for docente in docentes:
    if docente[0] != "D":
        print("Existe una irregularidad con los docentes:", docente)
        verificadora = False
        break
if verificadora:
    print("No existen irregularidades ni inconsistencias con los docentes.")
```

Para considerar las consistencias de esta variable es necesario verificar si existen registros diferentes del formato que poseen, es decir, serían inconsistentes aquellos datos que no empiezan con la letra "D", el resultado dio estableció que no existen inconsistencias.

Ilustración 573

Verificación de la variable area_asignatura

```
# Se visualizan los datos únicos de la variable
print("La variable cuenta con los siguientes datos únicos: ")
for dato in datos_procesados["area_asignatura"].unique():
    print(f"\t-{dato}")
```

La variable cuenta con los siguientes datos únicos:

```
-3.0
-7.0
-2.0
-6.0
-9.0
-11.0
-5.0
-4.0
-1.0
-10.0
```

Se puede observar que los datos se encuentran como numéricos de tipo flotante, es necesario cambiarlos a tipo "object" dado que no representan datos numéricos.

Ilustración 574

Cambio de tipo de dato de la variable area_asignatura y su verificación

Cambio de tipo de dato de la variable

```
# Se cambia primero a tipo entero para eliminar el punto decimal
datos_procesados["area_asignatura"] = datos_procesados["area_asignatura"].astype(int)

# Se transforma el tipo de "object"
datos_procesados["area_asignatura"] = datos_procesados["area_asignatura"].astype(object)
```

Verificación del cambio

```
datos_procesados["area_asignatura"].unique()
array([3, 7, 2, 6, 9, 11, 5, 4, 1, 10], dtype=object)
```

Se concluye que se realizó el cambio de tipo de dato para que tenga sentido la variable.

Ilustración 575

Verificación de la variable carrera

6.4.7. Verificación de la variable carrera

```
# Se visualizan los datos únicos de la variable
print("La variable cuenta con los siguientes datos únicos: ")
for dato in datos_procesados["carrera"].unique():
    print(f"\t-{dato}")
```

Se desplegaron todos los registros únicos del campo, se pueden visualizar en la siguientes ilustraciones.

Ilustración 576

Parte 1 del resultado de la verificación de la variable carrera

La variable cuenta con los siguientes datos únicos:

- Economía
- Arquitectura
- Antropología
- Sistemas de Información
- Terapia Física
- Sociología
- Psicología Clínica
- Psicología
- Enfermería
- Historia
- Lingüística Apl. Traducción
- Negocios Internacionales
- Sociología Política
- Fisioterapia
- Derecho
- Laboratorio Clínico
- Administración de Empresas
- Contabilidad y Auditoría
- Nutrición Humana
- Ciencias Biológicas
- Bioquímica Clínica
- Antropología: A. Sociocultural
- Microbiología
- Ingeniería Civil
- Teología
- Educación Inicial
- Cs Químicas Química Analít
- Lingüística
- Comunicación C. y Literatura
- Pedag Musical - Semipresencial
- Gestión Social y Desarrollo
- Com. Period. Pren. Rad y TV
- Química
- Artes Visuales
- Comunicación C. Organizacional
- Psicología Organizacional
- CC EE Educación Musical Semipr

Ilustración 577

Parte final del resultado de la verificación de la variable carrera

- Ecoturismo
- Gestión Social
- CC EE Educación Básica Semipre
- Relaciones Internacionales
- Tecnologías de la Información
- Educación Básica - Semipres.
- Finanzas
- Sociología R Internacionales
- Historia del Arte
- Biología
- Multilingüe Neg. R Internacion
- Diseño de Productos
- Diseño Gráfico
- Turismo
- Comunicación
- Diseño_Diseño de Productos
- Sociología Desarrollo
- CC EE Educación Inicial Semipr
- Psicología Educativa
- Mercadotecnia
- Educación Inicial Semipresenci
- Ing. de Sistemas y Computación
- Diseño Gráfico y Com. Visual
- Cs Geográficas Planif. Territ.
- Gestión Hotelera
- Nutrición y Dietética
- Ciencias Políticas
- Hospitalidad y Hotelería
- Filosofía
- Antropología - Arqueología
- Cs Geográficas Medio Ambiente
- CC EE Educación Inicial
- CC EE Música Leng y Movimien S
- Bioingeniería

En conclusión, no existen inconsistencias con la variable.

Ilustración 578

Verificación de la variable aprobado

```
# Se visualizan los datos únicos de la variable
print("La variable cuenta con los siguientes datos únicos: ")
for dato in datos_procesados["aprobado"].unique():
    print(f"\t-{dato}")
```

La variable cuenta con los siguientes datos únicos:

-S
-N

No existen inconsistencias.

Ilustración 579

Verificación de la variable creditos

```
# Se visualizan los datos únicos de la variable
print("La variable cuenta con los siguientes datos únicos: ")
for dato in datos_procesados["creditos"].unique():
    print(f"\t-{dato}")
```

La variable cuenta con los siguientes datos únicos:

-3.0
-4.5
-4.0
-0.0
-6.0
-2.0
-7.5
-8.0
-5.0
-10.0
-7.0
-1.0
-12.0
-12.5
-9.0
-11.0
-20.0
-19.0
-15.0
-3.75

No existen inconsistencias.

Vista del conjunto de datos

Ilustración 580

Vista del conjunto de datos

datos_procesados										
	etnia	nivel	segundas_matriculas	terceras_matriculas	facultad	docente	area_asignatura	carrera	aprobado	creditos
4	Mestizo	07	0.0	0.0	Facultad Economía	D177	3	Economía	S	3.0
5	Mestizo	07	0.0	0.0	Facultad Economía	D172	3	Economía	S	3.0
6	Mestizo	07	0.0	0.0	Facultad Economía	D158	3	Economía	S	3.0
7	Mestizo	07	0.0	0.0	Facultad Economía	D579	3	Economía	S	3.0
8	Mestizo	07	0.0	0.0	Facultad Economía	D507	3	Economía	S	3.0
...	...	...	...	...	...	...	...	...	...	...
38300	Mestizo	08	0.0	0.0	Facultad Arquitect Diseño y Artes	D44	7	Arquitectura	S	4.5
38301	Mulato	09	0.0	0.0	Facultad Arquitect Diseño y Artes	D839	7	Arquitectura	S	4.5
38302	Mulato	09	0.0	0.0	Facultad Arquitect Diseño y Artes	D336	7	Arquitectura	S	4.5
38303	Mulato	09	0.0	0.0	Facultad Arquitect Diseño y Artes	D329	7	Arquitectura	S	4.5
38304	Mulato	09	0.0	0.0	Facultad Arquitect Diseño y Artes	D280	7	Arquitectura	S	4.5

45826 rows × 10 columns

Vista de los tipos de datos del conjunto de datos

Ilustración 581

Vista de los tipos de datos del conjunto de datos

```
datos_procesados.dtypes

etnia          object
nivel          object
segundas_matriculas  float64
terceras_matriculas  float64
facultad       object
docente        object
area_asignatura object
carrera        object
aprobado       object
creditos       float64
dtype: object
```

Normalización y Estandarización de datos

En esta sección se realizará la normalización y estandarización de los datos, a continuación, se podrán visualizar los procesos que se llevaron a cabo con sus debidos comentarios:

Ilustración 582

Estandarización con StandardScaler

```
# Estandarización con StandardScaler
scaler = StandardScaler()
datos_procesados[
    ['segundas_matriculas',
     'terceras_matriculas',
     'creditos']] = scaler.fit_transform(datos_procesados[
    ['segundas_matriculas',
     'terceras_matriculas',
     'creditos']])
```

Este paso está diseñado para estandarizar las variables numéricas, este procedimiento es importante debido a que “permite asegurarse de que se tengan datos útiles y fácilmente enlazables y utilizables” (Ayuware, s. f.). En esta ocasión el uso de StandardScaler permite estandarizar la desviación estándar para que esta sea igual a 1.

Ilustración 583

Evaluación de la estandarización

Evaluación de la estandarización

```
datos_procesados.describe()
```

	segundas_matriculas	terceras_matriculas	creditos
count	4.582600e+04	4.582600e+04	4.582600e+04
mean	-4.569780e-14	-3.265408e-16	1.480604e-15
std	1.000011e+00	1.000011e+00	1.000011e+00
min	-3.036021e-01	-1.056228e-01	-2.536923e+00
25%	-3.036021e-01	-1.056228e-01	-2.913685e-01
50%	-3.036021e-01	-1.056228e-01	-2.913685e-01
75%	-3.036021e-01	-1.056228e-01	4.571496e-01
max	1.078382e+01	2.887802e+01	1.243344e+01

Como se puede observar, se efectuó correctamente la estandarización.

Ilustración 584

Normalización con MinMaxScaler

```
# Normalización
scaler = MinMaxScaler()
datos_procesados[
    ['segundas_matriculas',
     'terceras_matriculas',
     'creditos']] = scaler.fit_transform(datos_procesados[
    ['segundas_matriculas',
     'terceras_matriculas',
     'creditos']])
```

Este paso es similar al anterior, pero en lugar de estandarizar los datos, los normaliza para que todos los valores estén entre 0 y 1. El método MinMaxScaler funciona restando el valor mínimo de cada característica y luego dividiendo por el rango de esa característica.

Ilustración 585

Evaluación de la normalización

```
: datos_procesados.describe()
```

	segundas_matriculas	terceras_matriculas	creditos
count	45826.000000	45826.000000	45826.000000
mean	0.027383	0.003644	0.169463
std	0.090193	0.034503	0.066799
min	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.150000
50%	0.000000	0.000000	0.150000
75%	0.000000	0.000000	0.200000
max	1.000000	1.000000	1.000000

Como se puede observar, nuevamente se efectuó correctamente la normalización, pues, todos los valores se encuentran entre 0 y 1.

Ilustración 586

Codificación y evaluación de la variable aprobado

```
datos_procesados['aprobado'] = datos_procesados['aprobado'].map({'S': 1, 'N': 0})
```

Evaluación de la codificación

```
datos_procesados["aprobado"].unique()  
array([1, 0], dtype=int64)
```

Este paso convierte la variable 'aprobado', que asume que es una variable categórica con valores 'S' y 'N', en una variable binaria con valores 1 y 0. Esto es útil porque la mayoría de los modelos de machine learning requieren que las entradas sean numéricas.

Ilustración 587

Aplicación de One-Hot Encoding

```
# Se aplica la técnica de one-hot-encoding  
datos_procesados = pd.get_dummies(datos_procesados,  
                                   columns=["nivel",  
                                           "area_asignatura",  
                                           'etnia',  
                                           'facultad',  
                                           "carrera"])
```

Este paso convierte las variables categóricas en un formato que es adecuado para la mayoría de los algoritmos de machine learning, esta técnica conocida como One-Hot Encoding crea una nueva columna para cada valor único en la variable categórica, donde, cada una de estas nuevas columnas contiene 1 si el valor de la variable categórica para esa fila era igual a la columna y 0 en caso contrario.

Ilustración 588

Aplicación de Target Encoding para la variable docente

```
# Se aplica la técnica de target encoding  
encoder = ce.TargetEncoder(cols=['docente'])  
datos_procesados['docente'] = encoder.fit_transform(datos_procesados['docente'], datos_procesados['aprobado'])
```

Por último, este paso utiliza la técnica de codificación objetivo para la variable 'docente'. Este método reemplaza cada valor en 'docente' con la media de la variable objetivo (en este caso, 'aprobado') para las filas que tienen ese valor en 'docente'. Esto se hace con la ayuda de la biblioteca `category_encoders`. Se tomó la decisión de aplicar esta técnica sobre esta variable debido a la grande cantidad de datos diferentes de esta variable.

Ilustración 589

Revisión de las dimensiones

```
# Se muestra las dimensiones de los datos
print(f"Existen {datos_procesados.shape[0]} registros", end="")
print(f"y {datos_procesados.shape[1]} variables en el conjunto de datos unificado.")
Existen 45826 registrosy 118 variables en el conjunto de datos unificado.
```

Se puede visualizar que las dimensiones del conjunto de datos preparado para el modelo contienen 45826 registros y 118 variables, principalmente debido al One-Hot-Encoding.

4.3. Desarrollo de modelos de machine learning

4.3.1. Librerías requeridas

Las librerías que se requirieron fueron las que se encuentran debidamente comentadas en la siguiente ilustración:

```

# General para tratar con los datos
import pandas as pd
import numpy as np

# Sirve para dividir los datos en entrenamiento y prueba
from sklearn.model_selection import train_test_split

# Librerías para la creación de los modelos
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn import svm
from sklearn.neural_network import MLPClassifier

# Librerías para realizar la red neuronal con Keras
from keras.models import Sequential
from keras.utils import to_categorical
import tensorflow as tf
from tensorflow.keras.layers import Dense, Dropout, BatchNormalization
from tensorflow.keras.optimizers import Adam
from tensorflow.keras.initializers import HeNormal as He

# Librerías para validaciones
from sklearn.metrics import accuracy_score, recall_score, precision_score, f1_score, roc_auc_score

```

4.3.2. Selección de los modelos o algoritmos

Para esta investigación, se propone el uso de cinco modelos de machine learning distintos para abordar la tarea de predecir el rendimiento académico de los estudiantes universitarios. Los modelos seleccionados son:

- Regresión Logística
- Árbol de Decisión
- Random Forest
- Support Vector Machines (SVM)
- Redes Neuronales.

Estos modelos representan un rango de complejidad y flexibilidad, y se espera que al menos uno de ellos sea adecuado para este problema, sin embargo, la elección final del modelo dependerá de los resultados de la validación y la evaluación del modelo.

4.3.3. Entrenamiento y validación de los modelos

En este subcapítulo se realizará el proceso de entrenamiento y validación de los modelos de machine learning seleccionados para agendar la premisa del proyecto, predecir el rendimiento académico en términos de aprobación o reprobación de un estudiante, es decir, esclarecer que tan probable es que un estudiante reprueba una asignatura que está considerando cursar o va a comenzar a cursar. A continuación, se presentarán los modelos realizados como ilustraciones seguidas de un comentario, sin embargo, es necesario notar que las conclusiones de las validaciones de los modelos se realizarán en el siguiente capítulo.

Ilustración 590

División de los datos en entrenamiento y prueba

```
# Corresponde a los datos de entrada
X = datos_procesados.drop("aprobado", axis=1)
# Corresponde a los datos objetivo
y = datos_procesados["aprobado"]

# Se dividen los datos en entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

En este paso se realiza la división de los datos en entrenamiento y prueba.

Ilustración 591

Modelo de regresión logística

Creación del modelo

```
# Se crea el modelo
log_reg = LogisticRegression(max_iter=5000)
```

Entrenamiento del modelo

```
# Se entrena el modelo
log_reg.fit(X_train, y_train)

: LogisticRegression(max_iter=5000)
```

Predicción de las etiquetas para el conjunto de pruebas

```
y_pred = log_reg.predict(X_test)
```

Validaciones del modelo

```
accuracy = accuracy_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
roc_auc = roc_auc_score(y_test, y_pred)

print(f"Accuracy: {round(accuracy*100,2)}%")
print(f"Recall: {round(recall*100,2)}%")
print(f"Precision: {round(precision*100,2)}%")
print(f"F1-Score: {round(f1,5)}")
print(f"ROC AUC Score: {round(roc_auc,4)}")
```

```
Accuracy: 86.05%
Recall: 97.35%
Precision: 87.38%
F1-Score: 0.92098
ROC AUC Score: 0.6305
```

En este paso se realizó el primer modelo de predicción de rendimiento académico utilizando regresión logística de la librería de scikit-learn.

Tabla 41
Árbol de decisión

Creación del modelo

```
# Se crea el modelo  
tree = DecisionTreeClassifier()
```

Entrenamiento del modelo

```
# Se entrena el modelo  
tree.fit(X_train, y_train)  
:  
: DecisionTreeClassifier()
```

Predicción de las etiquetas para el conjunto de pruebas

```
y_pred = tree.predict(X_test)
```

Validaciones del modelo

```
accuracy = accuracy_score(y_test, y_pred)  
recall = recall_score(y_test, y_pred)  
precision = precision_score(y_test, y_pred)  
f1 = f1_score(y_test, y_pred)  
roc_auc = roc_auc_score(y_test, y_pred)  
  
print(f"Accuracy: {round(accuracy*100,2)}%")  
print(f"Recall: {round(recall*100,2)}%")  
print(f"Precision: {round(precision*100,2)}%")  
print(f"F1-Score: {round(f1,5)}")  
print(f"ROC AUC Score: {round(roc_auc,4)}")  
  
Accuracy: 94.96%  
Recall: 97.73%  
Precision: 96.29%  
F1-Score: 0.97005  
ROC AUC Score: 0.8933
```

En este paso se utiliza árboles de decisión para definir el segundo modelo de predicción de rendimiento académico de la librería de scikit-learn.

Ilustración 592

Random Forest

Creación del modelo

```
# Se crea el modelo
forest = RandomForestClassifier(n_estimators=100)
```

Entrenamiento del modelo

```
# Se entrena el modelo
forest.fit(X_train, y_train)
```

```
RandomForestClassifier()
```

Predicción de las etiquetas para el conjunto de pruebas

```
y_pred = forest.predict(X_test)
```

Validaciones del modelo

```
accuracy = accuracy_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
roc_auc = roc_auc_score(y_test, y_pred)

print(f"Accuracy: {round(accuracy*100,2)}%")
print(f"Recall: {round(recall*100,2)}%")
print(f"Precision: {round(precision*100,2)}%")
print(f"F1-Score: {round(f1,5)}")
print(f"ROC AUC Score: {round(roc_auc,4)}")
```

```
Accuracy: 95.61%
Recall: 98.73%
Precision: 96.12%
F1-Score: 0.9741
ROC AUC Score: 0.8927
```

En este paso se realizó el tercer modelo de predicción de rendimiento académico, en esta ocasión se utilizó Random Forest de la librería de scikit-learn.

Ilustración 593

Support Vector Machines (SVM)

Creación del modelo

```
# Se crea el modelo
svm_model = svm.SVC()
```

Entrenamiento del modelo

```
# Se entrena el modelo
svm_model.fit(X_train, y_train)
```

```
; svm_model =
```

Predicción de las etiquetas para el conjunto de pruebas

```
y_pred = svm_model.predict(X_test)
```

Validaciones del modelo

```
accuracy = accuracy_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
roc_auc = roc_auc_score(y_test, y_pred)

print(f"Accuracy: {round(accuracy*100,2)}%")
print(f"Recall: {round(recall*100,2)}%")
print(f"Precision: {round(precision*100,2)}%")
print(f"F1-Score: {round(f1,5)}")
print(f"ROC AUC Score: {round(roc_auc,4)}")
```

```
Accuracy: 86.74%
Recall: 98.86%
Precision: 87.03%
F1-Score: 0.9257
ROC AUC Score: 0.6208
```

El cuarto modelo empleado fue realizado mediante Support Vector Machines de la librería de scikit-learn.

Ilustración 594

Parte 1 de la primera propuesta de Red Neuronal

Creación del modelo

```
# Se crea el modelo
model = Sequential()
```

Se añaden capas

```
model.add(Dense(64, input_dim=117, activation='relu')) # Capa oculta 1: 64 neuronas
model.add(Dense(32, activation='relu')) # Capa oculta 2: 32 neuronas
model.add(Dense(1, activation='sigmoid')) # Capa de salida: 1 neurona, función de activación sigmoide
```

Se compila el modelo

```
model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
```

Ilustración 595

Parte 2 de la primera propuesta de Red Neuronal

Entrenamiento del modelo

```
# Se entrena el modelo
model.fit(X_train, y_train, validation_data=(X_test, y_test), epochs=10, batch_size=32)

Epoch 1/10
1146/1146 [=====] - 2s 2ms/step - loss: 0.3653 - accuracy: 0.8480 - val_loss: 0.3292 - val_accuracy: 0.8614
Epoch 2/10
1146/1146 [=====] - 2s 2ms/step - loss: 0.3326 - accuracy: 0.8600 - val_loss: 0.3094 - val_accuracy: 0.8690
Epoch 3/10
1146/1146 [=====] - 2s 2ms/step - loss: 0.2966 - accuracy: 0.8784 - val_loss: 0.2634 - val_accuracy: 0.8924
Epoch 4/10
1146/1146 [=====] - 2s 2ms/step - loss: 0.2416 - accuracy: 0.9041 - val_loss: 0.2180 - val_accuracy: 0.9194
Epoch 5/10
1146/1146 [=====] - 2s 2ms/step - loss: 0.1953 - accuracy: 0.9275 - val_loss: 0.1747 - val_accuracy: 0.9403
Epoch 6/10
1146/1146 [=====] - 2s 2ms/step - loss: 0.1624 - accuracy: 0.9423 - val_loss: 0.1511 - val_accuracy: 0.9499
Epoch 7/10
1146/1146 [=====] - 2s 2ms/step - loss: 0.1447 - accuracy: 0.9490 - val_loss: 0.1408 - val_accuracy: 0.9527
Epoch 8/10
1146/1146 [=====] - 2s 2ms/step - loss: 0.1362 - accuracy: 0.9518 - val_loss: 0.1331 - val_accuracy: 0.9518
Epoch 9/10
1146/1146 [=====] - 2s 2ms/step - loss: 0.1310 - accuracy: 0.9540 - val_loss: 0.1263 - val_accuracy: 0.9566
Epoch 10/10
1146/1146 [=====] - 2s 2ms/step - loss: 0.1266 - accuracy: 0.9558 - val_loss: 0.1201 - val_accuracy: 0.9576
<keras.callbacks.History at 0x2c68c4db430>
```

Ilustración 596

Parte final de la primera propuesta de Red Neuronal

Predicción de las etiquetas para el conjunto de pruebas

```
# Se obtienen las probabilidades de la clase positiva
y_pred_prob = model.predict(X_test)

# Se convierte las probabilidades a etiquetas de clase
y_pred = (y_pred_prob > 0.5).astype(int)

287/287 [=====] - 0s 1ms/step
```

Validaciones del modelo

```
# Se calculan las métricas de clasificación
accuracy = accuracy_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
roc_auc = roc_auc_score(y_test, y_pred_prob)

# Imprimir las métricas
print(f"Accuracy: {round(accuracy*100,2)}%")
print(f"Recall: {round(recall*100,2)}%")
print(f"Precision: {round(precision*100,2)}%")
print(f"F1-Score: {round(f1,5)}")
print(f"ROC AUC Score: {round(roc_auc,4)}")

Accuracy: 95.76%
Recall: 99.19%
Precision: 95.87%
F1-Score: 0.97503
ROC AUC Score: 0.9751
```

Estas tres últimas ilustraciones conforman a la primera propuesta de red neuronal realizado, en esta ocasión se utilizó keras para la creación de la propia red neuronal. La primera capa oculta consta de 64 neuronas y emplea la función de activación ReLU, de igual manera, la segunda capa oculta está compuesta por 32 neuronas que también utilizan la función de

activación ReLU, finalmente, la capa de salida está constituida por una única neurona que emplea la función de activación sigmoide.

Ilustración 597

Parte 1 de la segunda propuesta de red neuronal

Creación del modelo

```
# Se crea el modelo
model = Sequential()
```

Se añaden capas

```
model.add(Dense(128, input_dim=117, activation='relu')) # Capa oculta 1: 128 neuronas
model.add(Dense(64, activation='relu')) # Capa oculta 2: 64 neuronas
model.add(Dense(32, activation='relu')) # Capa oculta 3: 32 neuronas
model.add(Dense(16, activation='relu')) # Capa oculta 4: 16 neuronas
model.add(Dense(8, activation='relu')) # Capa oculta 5: 8 neuronas
model.add(Dense(1, activation='sigmoid'))
```

Se compila el modelo

```
model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
```

Ilustración 598

Parte 2 de la segunda propuesta de red neuronal

Entrenamiento del modelo

```
# Se entrena el modelo
model.fit(X_train, y_train, validation_data=(X_test, y_test), epochs=10, batch_size=32)

Epoch 1/10
1146/1146 [=====] - 3s 2ms/step - loss: 0.3688 - accuracy: 0.8489 - val_loss: 0.3284 - val_accuracy: 0.8634
Epoch 2/10
1146/1146 [=====] - 2s 2ms/step - loss: 0.3312 - accuracy: 0.8625 - val_loss: 0.3182 - val_accuracy: 0.8649
Epoch 3/10
1146/1146 [=====] - 2s 2ms/step - loss: 0.2564 - accuracy: 0.8970 - val_loss: 0.1640 - val_accuracy: 0.9412
Epoch 4/10
1146/1146 [=====] - 2s 2ms/step - loss: 0.1570 - accuracy: 0.9430 - val_loss: 0.1358 - val_accuracy: 0.9524
Epoch 5/10
1146/1146 [=====] - 2s 2ms/step - loss: 0.1359 - accuracy: 0.9516 - val_loss: 0.1203 - val_accuracy: 0.9583
Epoch 6/10
1146/1146 [=====] - 2s 2ms/step - loss: 0.1312 - accuracy: 0.9532 - val_loss: 0.1270 - val_accuracy: 0.9564
Epoch 7/10
1146/1146 [=====] - 2s 2ms/step - loss: 0.1280 - accuracy: 0.9548 - val_loss: 0.1228 - val_accuracy: 0.9563
Epoch 8/10
1146/1146 [=====] - 2s 2ms/step - loss: 0.1253 - accuracy: 0.9551 - val_loss: 0.1213 - val_accuracy: 0.9599
Epoch 9/10
1146/1146 [=====] - 2s 2ms/step - loss: 0.1211 - accuracy: 0.9569 - val_loss: 0.1211 - val_accuracy: 0.9573
Epoch 10/10
1146/1146 [=====] - 2s 2ms/step - loss: 0.1191 - accuracy: 0.9572 - val_loss: 0.1163 - val_accuracy: 0.9589
<keras.callbacks.History at 0x2c68c69b880>
```

Ilustración 599

Parte final de la segunda propuesta de red neuronal

Predicción de las etiquetas para el conjunto de pruebas

```
# Se obtienen las probabilidades de la clase positiva
y_pred_prob = model.predict(X_test)

# Se convierte las probabilidades a etiquetas de clase
y_pred = (y_pred_prob > 0.5).astype(int)

287/287 [=====] - 0s 865us/step
```

Validaciones del modelo

```
# Se calculan las métricas de clasificación
accuracy = accuracy_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
roc_auc = roc_auc_score(y_test, y_pred_prob)

# Imprimir las métricas
print(f"Accuracy: {round(accuracy*100,2)}%")
print(f"Recall: {round(recall*100,2)}%")
print(f"Precision: {round(precision*100,2)}%")
print(f"F1-Score: {round(f1,5)}")
print(f"ROC AUC Score: {round(roc_auc,4)}")

Accuracy: 95.89%
Recall: 99.39%
Precision: 95.84%
F1-Score: 0.97583
ROC AUC Score: 0.9774
```

Esta red neural consta de 5 capas, cada una definida de la siguiente manera:

- Capa oculta 1: 128 neuronas
- Capa oculta 2: 64 neuronas
- Capa oculta 3: 32 neuronas
- Capa oculta 4: 16 neuronas

- Capa oculta 5: 8 neuronas

Ilustración 600

Parte 1 de la tercera propuesta de red neuronal

Creación del modelo

```
# Se crea el modelo
model = Sequential()
```

Se añaden capas

```
model.add(Dense(512, input_dim=117, activation='relu')) # Capa oculta 1: 512 neuronas
model.add(Dense(256, activation='relu')) # Capa oculta 2: 256 neuronas
model.add(Dense(128, activation='relu')) # Capa oculta 3: 128 neuronas
model.add(Dense(64, activation='relu')) # Capa oculta 4: 64 neuronas
model.add(Dense(32, activation='relu')) # Capa oculta 5: 32 neuronas
model.add(Dense(1, activation='sigmoid'))
```

Se compila el modelo

```
model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
```

Ilustración 601

Parte 2 de la tercera propuesta de red neuronal

Entrenamiento del modelo

```
# Se entrena el modelo
model.fit(X_train, y_train, validation_data=(X_test, y_test), epochs=10, batch_size=32)

Epoch 1/10
1146/1146 [=====] - 5s 3ms/step - loss: 0.3581 - accuracy: 0.8517 - val_loss: 0.3240 - val_accuracy: 0.8628
Epoch 2/10
1146/1146 [=====] - 4s 3ms/step - loss: 0.2921 - accuracy: 0.8786 - val_loss: 0.2322 - val_accuracy: 0.9125
Epoch 3/10
1146/1146 [=====] - 3s 3ms/step - loss: 0.1964 - accuracy: 0.9285 - val_loss: 0.1725 - val_accuracy: 0.9398
Epoch 4/10
1146/1146 [=====] - 4s 4ms/step - loss: 0.1701 - accuracy: 0.9388 - val_loss: 0.1601 - val_accuracy: 0.9440
Epoch 5/10
1146/1146 [=====] - 4s 3ms/step - loss: 0.1579 - accuracy: 0.9412 - val_loss: 0.1579 - val_accuracy: 0.9414
Epoch 6/10
1146/1146 [=====] - 4s 3ms/step - loss: 0.1488 - accuracy: 0.9462 - val_loss: 0.1373 - val_accuracy: 0.9499
Epoch 7/10
1146/1146 [=====] - 4s 3ms/step - loss: 0.1359 - accuracy: 0.9513 - val_loss: 0.1439 - val_accuracy: 0.9516
Epoch 8/10
1146/1146 [=====] - 3s 3ms/step - loss: 0.1346 - accuracy: 0.9516 - val_loss: 0.1261 - val_accuracy: 0.9554
Epoch 9/10
1146/1146 [=====] - 4s 4ms/step - loss: 0.1249 - accuracy: 0.9556 - val_loss: 0.1269 - val_accuracy: 0.9523
Epoch 10/10
1146/1146 [=====] - 4s 3ms/step - loss: 0.1272 - accuracy: 0.9550 - val_loss: 0.1325 - val_accuracy: 0.9567
<keras.callbacks.History at 0x1d080297070>
```

Ilustración 602

Parte final de la tercera propuesta de red neuronal

Predicción de las etiquetas para el conjunto de pruebas

```
# Se obtienen las probabilidades de la clase positiva
y_pred_prob = model.predict(X_test)

# Se convierte las probabilidades a etiquetas de clase
y_pred = (y_pred_prob > 0.5).astype(int)

287/287 [=====] - 1s 1ms/step
```

Validaciones del modelo

```
# Se calculan las métricas de clasificación
accuracy = accuracy_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
roc_auc = roc_auc_score(y_test, y_pred_prob)

# Imprimir las métricas
print(f"Accuracy: {round(accuracy*100,2)}%")
print(f"Recall: {round(recall*100,2)}%")
print(f"Precision: {round(precision*100,2)}%")
print(f"F1-Score: {round(f1,5)}")
print(f"ROC AUC Score: {round(roc_auc,4)}")
```

```
Accuracy: 95.67%
Recall: 99.45%
Precision: 95.55%
F1-Score: 0.97459
ROC AUC Score: 0.9703
```

Esta red neuronal consta de 6 capas, cada una definida de la siguiente manera:

- Capa oculta 1: 512 neuronas
- Capa oculta 2: 256 neuronas
- Capa oculta 3: 128 neuronas
- Capa oculta 4: 64 neuronas
- Capa oculta 5: 32 neuronas

Ilustración 603

Parte 1 de la cuarta propuesta de red neuronal

Creación del modelo

```
# Se crea el modelo
model = Sequential()
```

Se añaden capas

```
model.add(Dense(1024, input_dim=117, activation='relu')) # Capa oculta 1: 1024 neuronas
model.add(Dense(512, activation='relu')) # Capa oculta 2: 512 neuronas
model.add(Dense(256, activation='relu')) # Capa oculta 3: 256 neuronas
model.add(Dense(128, activation='relu')) # Capa oculta 4: 128 neuronas
model.add(Dense(64, activation='relu')) # Capa oculta 5: 64 neuronas
model.add(Dense(32, activation='relu')) # Capa oculta 6: 32 neuronas
model.add(Dense(16, activation='relu')) # Capa oculta 7: 16 neuronas
model.add(Dense(1, activation='sigmoid'))
```

Se compila el modelo

```
model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
```

Ilustración 604

Parte 2 de la cuarta propuesta de red neuronal

Entrenamiento del modelo

```
# Se entrena el modelo
model.fit(X_train, y_train, validation_data=(X_test, y_test), epochs=10, batch_size=32)

Epoch 1/10
1146/1146 [=====] - 12s 9ms/step - loss: 0.3576 - accuracy: 0.8520 - val_loss: 0.3208 - val_accuracy: 0.8641
Epoch 2/10
1146/1146 [=====] - 10s 9ms/step - loss: 0.2792 - accuracy: 0.8890 - val_loss: 0.2125 - val_accuracy: 0.9217
Epoch 3/10
1146/1146 [=====] - 10s 9ms/step - loss: 0.1936 - accuracy: 0.9289 - val_loss: 0.1727 - val_accuracy: 0.9408
Epoch 4/10
1146/1146 [=====] - 10s 9ms/step - loss: 0.1685 - accuracy: 0.9387 - val_loss: 0.1647 - val_accuracy: 0.9401
Epoch 5/10
1146/1146 [=====] - 10s 9ms/step - loss: 0.1583 - accuracy: 0.9430 - val_loss: 0.1458 - val_accuracy: 0.9464
Epoch 6/10
1146/1146 [=====] - 10s 9ms/step - loss: 0.1477 - accuracy: 0.9476 - val_loss: 0.1472 - val_accuracy: 0.9503
Epoch 7/10
1146/1146 [=====] - 11s 9ms/step - loss: 0.1383 - accuracy: 0.9507 - val_loss: 0.1296 - val_accuracy: 0.9549
Epoch 8/10
1146/1146 [=====] - 10s 9ms/step - loss: 0.1337 - accuracy: 0.9518 - val_loss: 0.1325 - val_accuracy: 0.9531
Epoch 9/10
1146/1146 [=====] - 10s 9ms/step - loss: 0.1275 - accuracy: 0.9547 - val_loss: 0.1272 - val_accuracy: 0.9561
Epoch 10/10
1146/1146 [=====] - 11s 9ms/step - loss: 0.1236 - accuracy: 0.9558 - val_loss: 0.1244 - val_accuracy: 0.9581
]: <keras.callbacks.History at 0x1d0800baf40>
```

Ilustración 605

Parte final de la cuarta propuesta de red neuronal

Predicción de las etiquetas para el conjunto de pruebas

```
# Se obtienen Las probabilidades de la clase positiva
y_pred_prob = model.predict(X_test)

# Se convierte Las probabilidades a etiquetas de clase
y_pred = (y_pred_prob > 0.5).astype(int)
```

287/287 [=====] - 1s 3ms/step

Validaciones del modelo

```
# Se calculan Las métricas de clasificación
accuracy = accuracy_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
roc_auc = roc_auc_score(y_test, y_pred_prob)

# Imprimir Las métricas
print(f"Accuracy: {round(accuracy*100,2)}%")
print(f"Recall: {round(recall*100,2)}%")
print(f"Precision: {round(precision*100,2)}%")
print(f"F1-Score: {round(f1,5)}")
print(f"ROC AUC Score: {round(roc_auc,4)}")
```

Accuracy: 95.81%
Recall: 99.56%
Precision: 95.61%
F1-Score: 0.97543
ROC AUC Score: 0.9741

Esta red neuronal consta de 7 capa, cada una definida de la siguiente manera:

- Capa oculta 1: 1024 neuronas
- Capa oculta 2: 512 neuronas
- Capa oculta 3: 256 neuronas
- Capa oculta 4: 128 neuronas

- Capa oculta 5: 64 neuronas
- Capa oculta 6: 32 neuronas
- Capa oculta 7: 16 neuronas

Ilustración 606

Parte 1 de la quinta propuesta de red neuronal

Creación del modelo

```
# Se crea el modelo
model = Sequential()
```

Se añaden capas

```
# Capa de entrada
model.add(Dense(64, activation='relu', input_dim=117, kernel_initializer=He()))

# Capa oculta 1 con normalización por lotes y Dropout
model.add(BatchNormalization())
model.add(Dense(128, activation='relu', kernel_initializer=He()))
model.add(Dropout(0.5))

# Capa oculta 2 con normalización por lotes y Dropout
model.add(BatchNormalization())
model.add(Dense(128, activation='relu', kernel_initializer=He()))
model.add(Dropout(0.5))

# Capa de salida
model.add(Dense(1, activation='sigmoid'))
```

Se compila el modelo

```
model.compile(loss='binary_crossentropy', optimizer=Adam(learning_rate=0.001), metrics=['accuracy'])
```

Ilustración 607

Parte 2 de la quinta propuesta de red neuronal

Entrenamiento del modelo

```
# Se entrena el modelo
model.fit(X_train, y_train, validation_data=(X_test, y_test), epochs=50, batch_size=32)

Epoch 1/50
1146/1146 [=====] - 4s 3ms/step - loss: 0.4433 - accuracy: 0.8166 - val_loss: 0.3444 - val_accuracy: 0.8590
Epoch 2/50
1146/1146 [=====] - 3s 2ms/step - loss: 0.3645 - accuracy: 0.8474 - val_loss: 0.3303 - val_accuracy: 0.8638
Epoch 3/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.3533 - accuracy: 0.8515 - val_loss: 0.3227 - val_accuracy: 0.8673
Epoch 4/50
1146/1146 [=====] - 3s 2ms/step - loss: 0.3284 - accuracy: 0.8648 - val_loss: 0.2721 - val_accuracy: 0.8937
Epoch 5/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.2611 - accuracy: 0.8990 - val_loss: 0.2104 - val_accuracy: 0.9250
Epoch 6/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.2207 - accuracy: 0.9204 - val_loss: 0.1808 - val_accuracy: 0.9378
Epoch 7/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.2058 - accuracy: 0.9261 - val_loss: 0.1823 - val_accuracy: 0.9351
Epoch 8/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.1939 - accuracy: 0.9311 - val_loss: 0.1715 - val_accuracy: 0.9388
Epoch 9/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.1883 - accuracy: 0.9327 - val_loss: 0.1693 - val_accuracy: 0.9388
Epoch 10/50
1146/1146 [=====] - 3s 2ms/step - loss: 0.1799 - accuracy: 0.9364 - val_loss: 0.1651 - val_accuracy: 0.9428
Epoch 11/50
1146/1146 [=====] - 3s 2ms/step - loss: 0.1780 - accuracy: 0.9370 - val_loss: 0.1583 - val_accuracy: 0.9445
Epoch 12/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.1724 - accuracy: 0.9395 - val_loss: 0.1599 - val_accuracy: 0.9429
Epoch 13/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.1698 - accuracy: 0.9398 - val_loss: 0.1586 - val_accuracy: 0.9437
Epoch 14/50
1146/1146 [=====] - 3s 2ms/step - loss: 0.1678 - accuracy: 0.9409 - val_loss: 0.1480 - val_accuracy: 0.9482
Epoch 15/50
```

Ilustración 608

Parte 3 de la quinta propuesta de red neuronal

```
Epoch 35/50
1146/1146 [=====] - 3s 2ms/step - loss: 0.1374 - accuracy: 0.9508 - val_loss: 0.1319 - val_accuracy: 0.9553
Epoch 36/50
1146/1146 [=====] - 3s 2ms/step - loss: 0.1330 - accuracy: 0.9531 - val_loss: 0.1352 - val_accuracy: 0.9553
Epoch 37/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.1302 - accuracy: 0.9541 - val_loss: 0.1312 - val_accuracy: 0.9582
Epoch 38/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.1277 - accuracy: 0.9551 - val_loss: 0.1513 - val_accuracy: 0.9522
Epoch 39/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.1348 - accuracy: 0.9518 - val_loss: 0.1351 - val_accuracy: 0.9563
Epoch 40/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.1320 - accuracy: 0.9531 - val_loss: 0.1309 - val_accuracy: 0.9575
Epoch 41/50
1146/1146 [=====] - 3s 2ms/step - loss: 0.1276 - accuracy: 0.9547 - val_loss: 0.1487 - val_accuracy: 0.9512
Epoch 42/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.1262 - accuracy: 0.9549 - val_loss: 0.1287 - val_accuracy: 0.9570
Epoch 43/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.1295 - accuracy: 0.9545 - val_loss: 0.1259 - val_accuracy: 0.9573
Epoch 44/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.1283 - accuracy: 0.9544 - val_loss: 0.1264 - val_accuracy: 0.9576
Epoch 45/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.1266 - accuracy: 0.9552 - val_loss: 0.1337 - val_accuracy: 0.9583
Epoch 46/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.1255 - accuracy: 0.9558 - val_loss: 0.1275 - val_accuracy: 0.9568
Epoch 47/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.1276 - accuracy: 0.9545 - val_loss: 0.1370 - val_accuracy: 0.9568
Epoch 48/50
1146/1146 [=====] - 3s 2ms/step - loss: 0.1268 - accuracy: 0.9551 - val_loss: 0.1414 - val_accuracy: 0.9558
Epoch 49/50
1146/1146 [=====] - 3s 2ms/step - loss: 0.1262 - accuracy: 0.9550 - val_loss: 0.1221 - val_accuracy: 0.9591
Epoch 50/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.1227 - accuracy: 0.9564 - val_loss: 0.1309 - val_accuracy: 0.9590
<keras.callbacks.History at 0x2c68d9e47f0>
```

Ilustración 609

Parte final de la quinta propuesta de red neuronal

Predicción de las etiquetas para el conjunto de pruebas

```
.. # Se obtienen Las probabilidades de la clase positiva
y_pred_prob = model.predict(X_test)

# Se convierte Las probabilidades a etiquetas de clase
y_pred = (y_pred_prob > 0.5).astype(int)

287/287 [=====] - 1s 3ms/step
```

Validaciones del modelo

```
.. # Se calculan Las métricas de clasificación
accuracy = accuracy_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
roc_auc = roc_auc_score(y_test, y_pred_prob)

# Imprimir Las métricas
print(f"Accuracy: {round(accuracy*100,2)}%")
print(f"Recall: {round(recall*100,2)}%")
print(f"Precision: {round(precision*100,2)}%")
print(f"F1-Score: {round(f1,5)}")
print(f"ROC AUC Score: {round(roc_auc,4)}")

Accuracy: 95.9%
Recall: 99.66%
Precision: 95.61%
F1-Score: 0.97595
ROC AUC Score: 0.9728
```

Esta red neuronal consta de los siguientes puntos:

- Inicialización de He en todas las capas. Esta es una buena inicialización cuando se utiliza la función de activación ReLU.
- Capas de normalización por lotes después de la capa de entrada y las capas ocultas. Esto puede hacer que el entrenamiento sea más rápido y estable.
- Capas de Dropout después de las capas ocultas para prevenir el sobreajuste.

- Optimizador Adam, que es un buen punto de partida para muchos problemas. La tasa de aprendizaje es un hiperparámetro que podrías necesitar ajustar.

Ilustración 610

Parte 1 de la sexta propuesta de red neuronal

Creación del modelo

```
.. # Se crea el modelo
model = Sequential()
```

Se añaden capas

```
.. model.add(Dense(2048, input_dim=117, activation='relu')) # Capa oculta 1: 2048 neuronas
model.add(Dense(1024, activation='relu')) # Capa oculta 1: 1024 neuronas
model.add(Dropout(0.5))
model.add(Dense(512, activation='relu')) # Capa oculta 2: 512 neuronas
model.add(Dropout(0.5))
model.add(Dense(256, activation='relu')) # Capa oculta 3: 256 neuronas
model.add(Dropout(0.5))
model.add(Dense(128, activation='relu')) # Capa oculta 4: 128 neuronas
model.add(Dropout(0.5))
model.add(Dense(64, activation='relu')) # Capa oculta 5: 64 neuronas
model.add(Dropout(0.5))
model.add(Dense(32, activation='relu')) # Capa oculta 6: 32 neuronas
model.add(Dropout(0.5))
model.add(Dense(16, activation='relu')) # Capa oculta 7: 16 neuronas
model.add(Dropout(0.5))

model.add(Dense(1, activation='sigmoid'))
```

Se compila el modelo

```
.. model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
```

Ilustración 611

Parte 2 de la sexta propuesta de red neuronal

Entrenamiento del modelo

```
# Se entrena el modelo
model.fit(X_train, y_train, validation_data=(X_test, y_test), epochs=10, batch_size=32)

Epoch 1/10
1146/1146 [=====] - 36s 31ms/step - loss: 0.4331 - accuracy: 0.8298 - val_loss: 0.3824 - val_accuracy: 0.8353
Epoch 2/10
1146/1146 [=====] - 35s 31ms/step - loss: 0.3982 - accuracy: 0.8339 - val_loss: 0.4011 - val_accuracy: 0.8353
Epoch 3/10
1146/1146 [=====] - 37s 32ms/step - loss: 0.3665 - accuracy: 0.8339 - val_loss: 0.3419 - val_accuracy: 0.8353
Epoch 4/10
1146/1146 [=====] - 36s 31ms/step - loss: 0.3399 - accuracy: 0.8340 - val_loss: 0.3157 - val_accuracy: 0.8353
Epoch 5/10
1146/1146 [=====] - 35s 31ms/step - loss: 0.3188 - accuracy: 0.8340 - val_loss: 0.2924 - val_accuracy: 0.8353
Epoch 6/10
1146/1146 [=====] - 36s 32ms/step - loss: 0.3049 - accuracy: 0.8340 - val_loss: 0.2746 - val_accuracy: 0.8353
Epoch 7/10
1146/1146 [=====] - 35s 31ms/step - loss: 0.2911 - accuracy: 0.9042 - val_loss: 0.2553 - val_accuracy: 0.9133
Epoch 8/10
1146/1146 [=====] - 35s 31ms/step - loss: 0.2778 - accuracy: 0.9114 - val_loss: 0.2392 - val_accuracy: 0.9190
Epoch 9/10
1146/1146 [=====] - 36s 31ms/step - loss: 0.2688 - accuracy: 0.9128 - val_loss: 0.2481 - val_accuracy: 0.9092
Epoch 10/10
1146/1146 [=====] - 36s 31ms/step - loss: 0.2675 - accuracy: 0.9158 - val_loss: 0.2515 - val_accuracy: 0.9146
<keras.callbacks.History at 0x1d080424d60>
```

Ilustración 612

Parte final de la sexta propuesta de red neuronal

Predicción de las etiquetas para el conjunto de pruebas

```
# Se obtienen las probabilidades de la clase positiva
y_pred_prob = model.predict(X_test)

# Se convierte las probabilidades a etiquetas de clase
y_pred = (y_pred_prob > 0.5).astype(int)
```

287/287 [=====] - 2s 6ms/step

Validaciones del modelo

```
# Se calculan las métricas de clasificación
accuracy = accuracy_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
roc_auc = roc_auc_score(y_test, y_pred_prob)

# Imprimir las métricas
print(f"Accuracy: {round(accuracy*100,2)}%")
print(f"Recall: {round(recall*100,2)}%")
print(f"Precision: {round(precision*100,2)}%")
print(f"F1-Score: {round(f1,5)}")
print(f"ROC AUC Score: {round(roc_auc,4)}")
```

Accuracy: 91.46%

Recall: 93.42%

Precision: 96.25%

F1-Score: 0.9481

ROC AUC Score: 0.9412

Esta red neuronal consta de 8 capas, cada una definida de la siguiente manera:

- Capa oculta 1: 2048 neuronas
- Capa oculta 2: 1024 neuronas
- Capa oculta 3: 512 neuronas
- Capa oculta 4: 256 neuronas

- Capa oculta 5: 128 neuronas
- Capa oculta 6: 64 neuronas
- Capa oculta 7: 32 neuronas
- Capa oculta 8: 16 neuronas
- Capas de Dropout después de las capas ocultas para prevenir el sobreajuste.

Ilustración 613

Parte 1 de la séptima propuesta de red neuronal

Creación del modelo

```
# Se crea el modelo
model = Sequential()
```

Se añaden capas

```
model.add(Dense(2048, input_dim=117, activation='relu')) # Capa oculta 1: 2048 neuronas
model.add(Dense(1024, activation='relu')) # Capa oculta 1: 1024 neuronas
model.add(Dense(512, activation='relu')) # Capa oculta 2: 512 neuronas
model.add(Dense(256, activation='relu')) # Capa oculta 3: 256 neuronas
model.add(Dense(128, activation='relu')) # Capa oculta 4: 128 neuronas
model.add(Dense(64, activation='relu')) # Capa oculta 5: 64 neuronas
model.add(Dense(32, activation='relu')) # Capa oculta 6: 32 neuronas
model.add(Dense(16, activation='relu')) # Capa oculta 7: 16 neuronas

model.add(Dense(1, activation='sigmoid'))
```

Se compila el modelo

```
model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
```

Ilustración 614

Parte 2 de la séptima propuesta de red neuronal

Entrenamiento del modelo

```
# Se entrena el modelo
model.fit(X_train, y_train, validation_data=(X_test, y_test), epochs=10, batch_size=32)

Epoch 1/10
1146/1146 [=====] - 36s 31ms/step - loss: 0.3628 - accuracy: 0.8501 - val_loss: 0.3277 - val_accuracy: 0.8630
Epoch 2/10
1146/1146 [=====] - 37s 32ms/step - loss: 0.2978 - accuracy: 0.8787 - val_loss: 0.2123 - val_accuracy: 0.9161
Epoch 3/10
1146/1146 [=====] - 36s 32ms/step - loss: 0.1868 - accuracy: 0.9320 - val_loss: 0.1675 - val_accuracy: 0.9441
Epoch 4/10
1146/1146 [=====] - 35s 31ms/step - loss: 0.1561 - accuracy: 0.9447 - val_loss: 0.1472 - val_accuracy: 0.9500
Epoch 5/10
1146/1146 [=====] - 36s 31ms/step - loss: 0.1375 - accuracy: 0.9529 - val_loss: 0.1319 - val_accuracy: 0.9575
Epoch 6/10
1146/1146 [=====] - 37s 32ms/step - loss: 0.1308 - accuracy: 0.9530 - val_loss: 0.1365 - val_accuracy: 0.9520
Epoch 7/10
1146/1146 [=====] - 36s 32ms/step - loss: 0.1314 - accuracy: 0.9534 - val_loss: 0.1714 - val_accuracy: 0.9386
Epoch 8/10
1146/1146 [=====] - 36s 32ms/step - loss: 0.1272 - accuracy: 0.9549 - val_loss: 0.1271 - val_accuracy: 0.9569
Epoch 9/10
1146/1146 [=====] - 36s 31ms/step - loss: 0.1284 - accuracy: 0.9550 - val_loss: 0.1247 - val_accuracy: 0.9592
Epoch 10/10
1146/1146 [=====] - 37s 33ms/step - loss: 0.1193 - accuracy: 0.9570 - val_loss: 0.1275 - val_accuracy: 0.9570
<keras.callbacks.History at 0x2c6828289d0>
```

Ilustración 615

Parte final de la séptima propuesta de red neuronal

Predicción de las etiquetas para el conjunto de pruebas

```
# Se obtienen las probabilidades de la clase positiva
y_pred_prob = model.predict(X_test)

# Se convierte las probabilidades a etiquetas de clase
y_pred = (y_pred_prob > 0.5).astype(int)

287/287 [=====] - 2s 7ms/step
```

Validaciones del modelo

```
# Se calculan las métricas de clasificación
accuracy = accuracy_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
roc_auc = roc_auc_score(y_test, y_pred_prob)

# Imprimir las métricas
print(f"Accuracy: {round(accuracy*100,2)}%")
print(f"Recall: {round(recall*100,2)}%")
print(f"Precision: {round(precision*100,2)}%")
print(f"F1-Score: {round(f1,5)}")
print(f"ROC AUC Score: {round(roc_auc,4)}")
```

```
Accuracy: 95.7%
Recall: 99.02%
Precision: 95.96%
F1-Score: 0.97467
ROC AUC Score: 0.9763
```

Esta red neural consta de 8 capas, cada una definida de la siguiente manera:

- Capa oculta 1: 2048 neuronas
- Capa oculta 2: 1024 neuronas
- Capa oculta 3: 512 neuronas
- Capa oculta 4: 256 neuronas
- Capa oculta 5: 128 neuronas

- Capa oculta 6: 64 neuronas
- Capa oculta 7: 32 neuronas
- Capa oculta 8: 16 neuronas

4.4. Evaluación de los modelos de machine learning

De los modelos realizados, se pudo alcanzar la rescatar la siguiente tabla con los valores de las mediciones de todos los modelos creados:

Tabla 42
Tabla de evaluaciones de los modelos

Modelo	Accuracy	Recall	Precision	F1-score	ROC AUC Score
Regresión logística	86.05%	97.35%	87.38%	0.92	0.63
Árbol de decisión	94.96%	97.73%	96.29%	0.97	0.89
Random Forest	95.61%	98.73%	96.12%	0.97	0.89
Support Vector Machines (SVM)	86.74%	98.86%	87.03%	0.92	0.62
Primera propuesta de Red Neuronal	95.76%	99.19%	95.87%	0.97	0.97
Segunda propuesta de Red Neuronal	95.23%	99.58%	94.96%	0.97	0.97
Tercera propuesta de Red Neuronal	95.77%	99.45%	95.65%	0.97	0.97
Cuarta propuesta de Red Neuronal	95.81%	99.56%	95.61%	0.97	0.97
Quinta propuesta de Red Neuronal	95.9%	99.66%	95.61%	0.97	0.97
Sexta propuesta de Red Neuronal	91.46%	93.42%	96.25%	0.94	0.94
Séptima propuesta de Red Neuronal	95.7%	99.02%	95.96%	0.97	0.97

A partir de los datos de la tabla, se ha determinado que no es necesario realizar otro proceso de preparación ni de modelado, esto, debido a que los resultados son sin lugar a duda, sumamente positivos y se planea escoger un modelo de los creados hasta este momento; por lo que, a partir de los datos y una cuidadosa evaluación, se ha determinado que el modelo más acertado para el modelo de predicción de rendimiento académico, es la quinta propuesta de red neuronal, sin embargo, realmente se pudo considerar cualquiera de los demás modelos puesto que, teniendo en cuenta que se utilizaron datos reales, todos los modelos presentan una evaluación que abarca los criterios de éxito del modelo de machine learning planteado.

Sin embargo, en específico, el modelo fue seleccionado debido a sus puntajes de las evaluaciones, los cuales indican que de todas las clasificaciones que hace el modelo, aproximadamente el 95.9% son correctas, además, se indica que el modelo es muy eficiente al identificar positivos, y, de manera similar, indica que de todas las observaciones que el modelo predijo como positivas, aproximadamente el 95.61% realmente eran positivas; por otra parte, se indica que el modelo tiene un equilibrio muy bueno entre precisión y recall, finalmente, un puntaje de 0.97 en ROC AUC indica que el modelo tiene una capacidad excepcional para distinguir entre las categorías de los datos; sin embargo, como se ha mencionado, no se debería descartar a la ligera los demás modelos, puesto que, de igual manera tuvieron un rendimiento considerable y al ser todos los resultados bastante similares, la decisión final podría variar de acuerdo a las necesidades de la investigación, por lo que si usted estimado lector, tiene tal vez otra perspectiva con respecto a un mejor modelo, es invitado a experimentarlo y usarlo a gusto.

CAPÍTULO V: RESULTADO

5. Resultado

5.1. Descripción de los conjuntos de datos iniciales

Los datos se obtuvieron tras un acuerdo con la Dirección de Bienestar Estudiantil de la Pontificia Universidad Católica del Ecuador, por lo que, al ser datos reales de la comunidad universitaria se tuvo que firmar un acuerdo de confidencialidad respectivo. A partir de este acuerdo se obtuvieron registros correspondientes a 3 diferentes reportes, los reportes obtenidos fueron:

- Reporte de Asignaturas perdidas por periodo y reincidencia
- Reporte de Calificaciones por carrera
- Reporte de detalle de estudiantes inscritos con beca

A partir de estos reportes, por cada uno se obtuvieron 4 periodos académicos, estos fueron:

- Periodo académico 2021-01
- Periodo académico 2021-02
- Periodo académico 2022-01
- Periodo académico 2021-02

De manera general, todos los reportes correspondían a Pregrado y al campus de la Pontificia Universidad Católica del Ecuador en Quito, sin embargo, en específico, cada reporte consistía en lo siguiente:

- **Reporte de asignaturas perdidas por periodo y reincidencia:** Este reporte contiene registros de los estudiantes que perdieron asignaturas en un determinado periodo, por

ejemplo, para el reporte del periodo 2021-01 se contiene información de los estudiantes que perdieron sus asignaturas en el semestre 2021-01.

- **Reporte de calificaciones por carrera:** Este reporte contiene registros de las calificaciones de cada estudiante tanto parciales como finales, así como también se tiene presente las carreras de los estudiantes, asignaturas y profesores.
- **Reporte de detalle de estudiantes inscritos con beca:** Este reporte contiene registros de los estudiantes que poseen becas, discapacidades, cantidad de asignaturas con segunda o tercera matrícula, e información de pago de estos.

5.2. Resultados de la exploración de datos

Para empezar, la exploración de datos permitió identificar variables relevantes, irrelevantes, necesarias para cumplir ciertos procesos (como al unificación de los conjuntos de datos), y la variable objetivo; esto implica que la exploración de datos permitió la reducción de todas las variables que ostentaban los diferentes reportes, dando como resultado la siguiente tabla de variables y descripciones:

Tabla 43
Descripción de Variables Resultantes

Variable	Descripción
etnia	Etnia a la que pertenece un estudiante.
nivel	Nivel académico de un estudiante, en términos del semestre en el que se encuentra este.

segundas_matriculas	Número de materias en segunda matrícula de un estudiante.
terceras_matriculas	Número de materias en tercera matrícula de un estudiante.
facultad	Nombre de la facultad a la que pertenece un estudiante.
docente	Identificador único de un docente en el sistema de gestión académica.
area_asignatura	Área de la asignatura en el que se encuentra inscrito un estudiante.
carrera	Nombre de la carrera en el que se encuentra inscrito un estudiante.
aprobado	Indicador si un estudiante aprobó o reprobó el curso.
creditos	Número de créditos asignados a la asignatura de un estudiante.

En otras palabras, se logró identificar certeramente las variables que tienen influencia sobre el rendimiento académico en los estudiantes de la Pontificia Universidad Católica del Ecuador, por lo que, en conclusión, a partir de la exploración de los datos, se pudo formular una hipótesis para el modelo de predicción de rendimiento académico, esta, permitió resultados

sumamente precisos con respecto a sus validaciones y evaluaciones, es decir, cumplió con las expectativas de los criterios de éxito declarados al comienzo.

La hipótesis del modelo de predicción del rendimiento académico planteado fue:

Las variables referidas a la facultad a la que pertenece un estudiante, su carrera, su nivel académico, el área de asignatura que cursará o está considerando cursar, la cantidad de créditos que equivale la asignatura, el docente que imparte dicha asignatura, la etnia, su cantidad de asignaturas con segunda y tercera matrícula influyen directamente en el rendimiento académico de los estudiantes universitarios en la Pontificia Universidad Católica del Ecuador.

5.3. Resultados del modelado del rendimiento académico

A partir de todos los procesos llevados a cabo en el presente trabajo de titulación, se crearon y entrenaron 12 diferentes modelos de predicción del rendimiento académico, de los cuales, de manera general, todos sus resultados, en términos de sus evaluaciones, fueron bastante positivos; por lo que se puede concretar que los criterios de éxito remarcados al comienzo de la investigación fueron cumplidos satisfactoriamente.

Adicionalmente, los resultados de los modelos de predicción de rendimiento académico realizados pueden ser apreciados en la siguiente tabla:

Tabla 44
Resultados de los modelos realizados

Modelo	Accuracy	Recall	Precision	F1-score	ROC AUC Score
Regresión logística	86.05%	97.35%	87.38%	0.92	0.63
Árbol de decisión	94.96%	97.73%	96.29%	0.97	0.89
Random Forest	95.61%	98.73%	96.12%	0.97	0.89

Support Vector Machines (SVM)	86.74%	98.86%	87.03%	0.92	0.62
Primera propuesta de Red Neuronal	95.76%	99.19%	95.87%	0.97	0.97
Segunda propuesta de Red Neuronal	95.23%	99.58%	94.96%	0.97	0.97
Tercera propuesta de Red Neuronal	95.77%	99.45%	95.65%	0.97	0.97
Cuarta propuesta de Red Neuronal	95.81%	99.56%	95.61%	0.97	0.97
Quinta propuesta de Red Neuronal	95.9%	99.66%	95.61%	0.97	0.97
Sexta propuesta de Red Neuronal	91.46%	93.42%	96.25%	0.94	0.94
Séptima propuesta de Red Neuronal	95.7%	99.02%	95.96%	0.97	0.97

5.4. Comparación de modelos

A partir de los datos visibles en la tabla de la comparativa de modelos, se pudo determinar que la quinta propuesta de red neuronal es la más adecuada para el objetivo de predecir el rendimiento académico universitario de los estudiante de la Pontificia Universidad Católica del Ecuador, debido a que, de todas las clasificaciones que hace el modelo, aproximadamente el 95.9% son correctas, además, se indica que el modelo es muy eficiente al identificar positivos, y, de manera similar, indica que de todas las observaciones que el modelo predijo como positivas, aproximadamente el 95.61% realmente eran positivas; por otra parte, se indica que el modelo tiene un equilibrio muy bueno entre precisión y recall, finalmente, un puntaje de 0.97 en ROC AUC indica que el modelo tiene una capacidad excepcional para distinguir entre las categorías de los datos.

Sin embargo, a pesar de que se remarque que el modelo seleccionado como el más acertado fue la quinta propuesta de red neuronal, las demás propuestas presentan calificaciones aceptables, sobre todo cuando se considera que se realizó el modelo con datos reales, esto implica que la decisión con respecto al modelo más asertivo puede darse en cuanto a una selección múltiple de los modelos realizados. De todas maneras, para esta investigación se determinó la quinta propuesta de red neuronal para el despliegue del modelo de predicción del rendimiento académico.

CAPÍTULO VI: DISCUSIÓN

6. Discusión

6.1. Interpretación de resultados

En el presente capítulo se detallarán resultados relevantes hallados a lo largo de la investigación de manera concisa y precisa.

6.1.1. Una posible relación entre la modalidad virtual, el descenso del rendimiento académico y el decrecimiento de la cantidad de estudiantes matriculados

Durante los análisis realizados se pudo determinar que progresivamente el rendimiento general de los estudiantes de la Pontificia Universidad Católica del Ecuador ha ido decreciendo, esto podría estar relacionado a que los primeros periodos de los que se obtuvo información para esta investigación, consistían a los periodos que estuvieron en modalidad virtual debido a la pandemia, por lo que, es posible que la transición de modalidad virtual a presencial haya tenido un impacto significativo en el rendimiento académico de los estudiantes, puesto que, de todos los estudiantes, existe un gran porcentaje de estos que para el periodo 2022-01 fue la primera vez que conocieron la universidad físicamente en términos académicos, después de haber cursado varios semestres en virtualidad.

Por otra parte, también podría deberse a problemas en la calidad de educación virtual de su momento, sin embargo, como se ha demostrado repetidamente en esta investigación, no es posible definir el rendimiento académico en términos de que solo un factor influya en como un estudiante se desempeñará, por lo que, tal vez la problemática no se encuentra exclusivamente ligada a como la universidad gestiona la educación, sino a factores relaciones con la dedicación de los estudiantes, pues, a parte de los factores que pueden influenciar externamente en un estudiante, también existe el factor interno, el cual, por obra del estudiante permite que se

decrezca su rendimiento. Por lo que, a partir de lo mencionado, es posible que, durante la época de la modalidad virtual, dada la manera en la que se conllevó la educación, esta haya resultado muy inferior en términos de exigencia con respecto a la modalidad presencial posterior, lo que muy posiblemente haya repercutido en estudiantes que no han estado acostumbrados a la intensidad de los semestres presenciales.

Teniendo en cuenta lo mencionado hasta el momento, también es importante comentar que de los análisis realizados, también se puede observar un patrón de comportamiento en el cual progresivamente se ha ido disminuyendo la cantidad de estudiantes matriculados por periodo, es decir, dado los datos obtenidos por la dirección de estudiantes de la universidad, se presenció como cada semestre el aumento de estudiantes no es proporcional al decrecimiento de los mismos, en otras palabras, existen más estudiantes que cesan de estudiar en un periodo que aquellos que comienzan a hacerlo; no obstante, si bien este acontecimiento puede deberse a varios factores ajenos al rendimiento académico, no se debería dejar de lado la idea de que talvez el decrecimiento del rendimiento académico general de la universidad y el descenso de la cantidad de estudiantes matriculados en la misma, este relacionado.

6.1.2. *Historia, becas socioeconómicas y sus beneficiarios*

A partir de la investigación realizada se pudo determinar una característica común entre los estudiantes quienes fueron partícipes de una beca socioeconómica dentro de los periodos 2021-01 hasta el 2022-02, el comportamiento destacó que la mayor parte de los estudiantes acreedores de este beneficio provenían de colegios particulares religiosos, por lo que podría sugerir que si un estudiante formó parte de un colegio particular religioso en su etapa escolar, podría tender a tener una mayor probabilidad de ser beneficiario de una beca socioeconómica.

Sin embargo, el punto más resaltante de los resultados de la investigación con respecto a las becas socioeconómicas que poseían los estudiantes, recae en que es probable que no haya sido notable una relación de esta con respecto al rendimiento académico, debido a que durante la época de la pandemia, a todos los estudiantes que ingresaban se les asignaba automáticamente la categoría C, en otras palabras, dado que los datos obtenidos fueron a partir de la pandemia y que la asociación de estas becas con los estudiantes dejó de remarcar los contextos socioculturales de cada individuo, cesó la relación del rendimiento académico con la beca socioeconómica; esto sugiere que para analizar a fondo si existe una posible relación entre las becas socioeconómicas y el rendimiento académico, es necesario contar con mayor cantidad de periodos académicos.

6.2. Contribuciones y limitaciones del estudio

De manera general las limitaciones del estudio son principalmente aquellas dadas por los datos, si bien se puede gestionar un extenso análisis de un montón de datos, en el caso que estos no representen la realidad, los análisis pueden resultar inservibles dejando meses de trabajo de investigación en la basura. De todas maneras, por fortuna este parece no ser el caso, debido a que los datos fueron recopilados a través de la dirección de estudiantes de la Pontificia Universidad Católica del Ecuador, por lo que, fueron datos que expresan la realidad que viven los estudiantes durante los diferentes periodos académicos.

Adicionalmente, una de las mayores limitaciones que se presentaron fue la cantidad de información, esto con respecto al límite de información que existe, es decir, esta investigación fue bastante extensa, sin embargo, pudo serlo más; el motivo por el que se evitó extender más la investigación fue debido a que este trabajo debe ser revisado por tutores asignados y no es prudente enviar una inmensa cantidad de palabras cuando los mismos asesores tienen otros trabajos que revisar de igual manera.

Por otra parte, con respecto a las contribuciones de este proyecto, se puede recalcar un extenso análisis que fue detallado en la sección anterior, justamente en el subcapítulo antes de este, además, se ha compartido el conjunto de datos procesado acompañado de este trabajo de titulación con la finalidad de que futuros investigadores puedan usar estos datos y ahorrarse una considerable cantidad de tiempo.

6.3. Implicaciones para investigación futura

Para empezar, una de las mayores implicaciones que se deberían tener para investigaciones futuras es contar con una mayor cantidad de datos y considerar las situaciones históricas, geopolíticas y culturales de aquellos determinados periodos, por ejemplo, los datos utilizados para esta investigación fueron a partir del fin de la pandemia, por lo que muy posiblemente fueron sesgados por ella y las implicaciones que afectaron a tanto estudiantes, profesores y la propia universidad.

Por otra parte, se invita a los investigadores que por algún azar revisen esta investigación, que prueben enfoques e hipótesis para el modelo de rendimiento académico diferentes, pero, antes de ello, están cordialmente invitados a aprovechar los datos que se han compartido con ustedes, pues, al ser datos reales y seleccionados tras un extenso análisis de relaciones, el conjunto de datos llamado “datos_para_el_investigador.csv” contiene información suficiente para realizar modelos de machine learning con diversidad de enfoques.

CAPÍTULO VII: DESPLIEGUE DEL MODELO

7. Despliegue del modelo

7.1. Despliegue del modelo de machine learning escogido

La fase de despliegue corresponde al último paso de la metodología CRISP-DM, en esta fase, tal como se mencionó en el capítulo 2, se realizará el despliegue del modelo más acertado realizado y evaluado en el capítulo 4, por lo que, para cumplir con este objetivo se prevé realizar un aplicativo web simple que permita la interacción con algún usuario, el mismo, podrá observar detalles generales del proyecto, pero, en específico, el usuario podrá ingresar datos de un estudiante y predecir su rendimiento académico en un asignatura en términos de aprobación o reprobación.

Para lograr el desarrollo del aplicativo web con lo mencionado, es primordial primero volver a ejecutar nuevamente la creación y entrenamiento del modelo de la quinta propuesta de red neuronal, esto, con la finalidad de exportar el modelo para su uso en el despliegue, además, se aprovechará para exportar las opciones para las áreas de asignatura, facultades, carreras, etc. En otras palabras, lo que se realizará antes de la programación del aplicativo web, es volver a crear, entrenar y evaluar el modelo para exportarlo como un objeto y usarlo indefinidamente, asimismo, se procederá a exportar en archivos las categorías existentes en cada variable, esto, con la finalidad de permitir al usuario que manipule el aplicativo web ingresar los datos del estudiante conforme a los datos existentes de cada categoría usados para la creación del modelo.

Por lo que, se comenzará con el primer objetivo de esta fase de despliegue, es decir, exportar el modelo de la quinta propuesta de red neuronal realizada, para ello, como se ha mencionado se deberá realizar su creación, entrenamiento y evaluación correspondiente, no

obstante, es necesario mencionar que la fase de preparación de los datos es obviada debido a que será exactamente la misma que la realizada en el capítulo 4.

Ilustración 616

Creación de la quinta propuesta de red neuronal

Creación del modelo

```
# Se crea el modelo
model = Sequential()
```

Se añaden capas

```
# Capa de entrada
model.add(Dense(64, activation='relu', input_dim=117, kernel_initializer=He()))

# Capa oculta 1 con normalización por Lotes y Dropout
model.add(BatchNormalization())
model.add(Dense(128, activation='relu', kernel_initializer=He()))
model.add(Dropout(0.5))

# Capa oculta 2 con normalización por Lotes y Dropout
model.add(BatchNormalization())
model.add(Dense(128, activation='relu', kernel_initializer=He()))
model.add(Dropout(0.5))

# Capa de salida
model.add(Dense(1, activation='sigmoid'))
```

Se compila el modelo

```
model.compile(loss='binary_crossentropy', optimizer=Adam(learning_rate=0.001), metrics=['accuracy'])
```

Ilustración 617

Primera parte del entrenamiento de la quinta propuesta de red neuronal

Entrenamiento del modelo

```
# Se entrena el modelo
model.fit(X_train, y_train, validation_data=(X_test, y_test), epochs=50, batch_size=32)

Epoch 1/50
1146/1146 [=====] - 5s 3ms/step - loss: 0.4254 - accuracy: 0.8273 - val_loss: 0.3453 - val_accuracy: 0.8586
Epoch 2/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.3627 - accuracy: 0.8495 - val_loss: 0.3349 - val_accuracy: 0.8636
Epoch 3/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.3446 - accuracy: 0.8560 - val_loss: 0.3093 - val_accuracy: 0.8732
Epoch 4/50
1146/1146 [=====] - 3s 2ms/step - loss: 0.2948 - accuracy: 0.8799 - val_loss: 0.2404 - val_accuracy: 0.9058
Epoch 5/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.2495 - accuracy: 0.9020 - val_loss: 0.2019 - val_accuracy: 0.9274
Epoch 6/50
1146/1146 [=====] - 3s 2ms/step - loss: 0.2142 - accuracy: 0.9207 - val_loss: 0.1837 - val_accuracy: 0.9337
Epoch 7/50
1146/1146 [=====] - 3s 2ms/step - loss: 0.1896 - accuracy: 0.9309 - val_loss: 0.1823 - val_accuracy: 0.9383
Epoch 8/50
1146/1146 [=====] - 3s 2ms/step - loss: 0.1827 - accuracy: 0.9352 - val_loss: 0.1636 - val_accuracy: 0.9453
Epoch 9/50
1146/1146 [=====] - 3s 2ms/step - loss: 0.1748 - accuracy: 0.9379 - val_loss: 0.1839 - val_accuracy: 0.9364
Epoch 10/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.1757 - accuracy: 0.9382 - val_loss: 0.1683 - val_accuracy: 0.9411
Epoch 11/50
1146/1146 [=====] - 3s 2ms/step - loss: 0.1668 - accuracy: 0.9400 - val_loss: 0.1591 - val_accuracy: 0.9479
Epoch 12/50
1146/1146 [=====] - 3s 2ms/step - loss: 0.1584 - accuracy: 0.9447 - val_loss: 0.1661 - val_accuracy: 0.9459
Epoch 13/50
1146/1146 [=====] - 3s 2ms/step - loss: 0.1618 - accuracy: 0.9418 - val_loss: 0.1779 - val_accuracy: 0.9431
Epoch 14/50
1146/1146 [=====] - 3s 2ms/step - loss: 0.1642 - accuracy: 0.9406 - val_loss: 0.1670 - val_accuracy: 0.9445
Epoch 15/50
1146/1146 [=====] - 3s 2ms/step - loss: 0.1592 - accuracy: 0.9438 - val_loss: 0.1496 - val_accuracy: 0.9483
Epoch 16/50
1146/1146 [=====] - 3s 2ms/step - loss: 0.1506 - accuracy: 0.9458 - val_loss: 0.1404 - val_accuracy: 0.9521
Epoch 17/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.1611 - accuracy: 0.9408 - val_loss: 0.1500 - val_accuracy: 0.9475
Epoch 18/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.1557 - accuracy: 0.9431 - val_loss: 0.1452 - val_accuracy: 0.9506
Epoch 19/50
1146/1146 [=====] - 3s 2ms/step - loss: 0.1477 - accuracy: 0.9465 - val_loss: 0.1549 - val_accuracy: 0.9458
Epoch 20/50
1146/1146 [=====] - 3s 2ms/step - loss: 0.1558 - accuracy: 0.9424 - val_loss: 0.1406 - val_accuracy: 0.9520
```

Ilustración 618

Parte final del entrenamiento de la quinta propuesta de red neuronal

```
1146/1146 [=====] - 3s 3ms/step - loss: 0.1351 - accuracy: 0.9517 - val_loss: 0.1281 - val_accuracy: 0.9570
Epoch 30/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.1319 - accuracy: 0.9538 - val_loss: 0.1235 - val_accuracy: 0.9579
Epoch 31/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.1309 - accuracy: 0.9540 - val_loss: 0.1188 - val_accuracy: 0.9578
Epoch 32/50
1146/1146 [=====] - 4s 3ms/step - loss: 0.1319 - accuracy: 0.9534 - val_loss: 0.1262 - val_accuracy: 0.9563
Epoch 33/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.1324 - accuracy: 0.9530 - val_loss: 0.1254 - val_accuracy: 0.9582
Epoch 34/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.1346 - accuracy: 0.9534 - val_loss: 0.1233 - val_accuracy: 0.9569
Epoch 35/50
1146/1146 [=====] - 3s 2ms/step - loss: 0.1301 - accuracy: 0.9536 - val_loss: 0.1224 - val_accuracy: 0.9578
Epoch 36/50
1146/1146 [=====] - 3s 2ms/step - loss: 0.1318 - accuracy: 0.9531 - val_loss: 0.1220 - val_accuracy: 0.9578
Epoch 37/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.1299 - accuracy: 0.9540 - val_loss: 0.1203 - val_accuracy: 0.9567
Epoch 38/50
1146/1146 [=====] - 3s 2ms/step - loss: 0.1274 - accuracy: 0.9544 - val_loss: 0.1154 - val_accuracy: 0.9591
Epoch 39/50
1146/1146 [=====] - 3s 2ms/step - loss: 0.1277 - accuracy: 0.9551 - val_loss: 0.1186 - val_accuracy: 0.9570
Epoch 40/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.1249 - accuracy: 0.9548 - val_loss: 0.1137 - val_accuracy: 0.9596
Epoch 41/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.1282 - accuracy: 0.9552 - val_loss: 0.1209 - val_accuracy: 0.9582
Epoch 42/50
1146/1146 [=====] - 4s 3ms/step - loss: 0.1288 - accuracy: 0.9545 - val_loss: 0.1147 - val_accuracy: 0.9597
Epoch 43/50
1146/1146 [=====] - 4s 3ms/step - loss: 0.1265 - accuracy: 0.9557 - val_loss: 0.1250 - val_accuracy: 0.9572
Epoch 44/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.1243 - accuracy: 0.9552 - val_loss: 0.1161 - val_accuracy: 0.9596
Epoch 45/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.1230 - accuracy: 0.9559 - val_loss: 0.1191 - val_accuracy: 0.9592
Epoch 46/50
1146/1146 [=====] - 3s 2ms/step - loss: 0.1232 - accuracy: 0.9568 - val_loss: 0.1272 - val_accuracy: 0.9572
Epoch 47/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.1215 - accuracy: 0.9567 - val_loss: 0.1257 - val_accuracy: 0.9559
Epoch 48/50
1146/1146 [=====] - 3s 3ms/step - loss: 0.1232 - accuracy: 0.9564 - val_loss: 0.1243 - val_accuracy: 0.9580
Epoch 49/50
1146/1146 [=====] - 3s 2ms/step - loss: 0.1222 - accuracy: 0.9565 - val_loss: 0.1215 - val_accuracy: 0.9578
Epoch 50/50
1146/1146 [=====] - 3s 2ms/step - loss: 0.1226 - accuracy: 0.9564 - val_loss: 0.1208 - val_accuracy: 0.9581
<keras.callbacks.History at 0x2c6918558e0>
```

Ilustración 619

Evaluación de la quinta propuesta de red neuronal

Predicción de las etiquetas para el conjunto de pruebas

```
... # Se obtienen las probabilidades de la clase positiva
y_pred_prob = model.predict(X_test)

# Se convierte las probabilidades a etiquetas de clase
y_pred = (y_pred_prob > 0.5).astype(int)

287/287 [=====] - 0s 966us/step
```

Validaciones del modelo

```
... # Se calculan las métricas de clasificación
accuracy = accuracy_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
roc_auc = roc_auc_score(y_test, y_pred_prob)

# Imprimir las métricas
print(f"Accuracy: {round(accuracy*100,2)}%")
print(f"Recall: {round(recall*100,2)}%")
print(f"Precision: {round(precision*100,2)}%")
print(f"F1-Score: {round(f1,5)}")
print(f"ROC AUC Score: {round(roc_auc,4)}")

Accuracy: 95.81%
Recall: 99.67%
Precision: 95.51%
F1-Score: 0.97546
ROC AUC Score: 0.9775
```

En este momento, se ha finalizado el proceso de la creación, entrenamiento y validación de la quinta propuesta de red neuronal, no obstante, es necesario remarcar que es imperativo que durante la validación del modelo, se visualicen resultados similares a los detallados en el capítulo 4, por lo que, en esta ocasión, al observarse un comportamiento bastante similar y en algunos casos superior a los registrados en su desarrollo anterior, se puede proceder a exportar el modelo para su uso en el despliegue de este, es decir, en el aplicativo web planificada. Para realizar esta

acción se utilizará el método “save” sobre el modelo creado, este método es definido por Keras para permitir la exportación de modelos en formatos como “h5” o archivo de datos en formato de datos jerárquicos HDF, por esta razón, el modelo exportado se llamará de la siguiente manera: “modelo_tesis.h5”; el código para realizar esta acción se puede visualizar a continuación:

Ilustración 620

Exportación del modelo de la quinta propuesta de red neuronal

```
model.save("modelo_tesis.h5")
```

Una vez se exporto el modelo, es momento de realizar el segundo objetivo de este proceso, es decir, la exportación de las categorías existentes por cada variable teniendo en cuenta sus codificaciones, esto, con la finalidad de permitir al usuario que interactúe con el aplicativo web el poder manipular las predicciones de los estudiantes con datos que fueron utilizados para el desarrollo del modelo. Por lo que, para llevar a cabo esta tarea es necesario tener en cuenta la forma en la que se codificaron estos datos.

Para empezar, es necesario volver a realizar los procesos de preparación de los datos hasta antes de la estandarización, normalización y codificaciones extras, debido a que se debe crear una copia de los datos en la que no se realizarán ninguno de los procedimientos subsecuentes.

Ilustración 621

Creación de una copia de los datos antes de la estandarización y normalización

```
temp = datos_procesados.copy()
```

Una vez realizado este paso, se ejecuta la estandarización sobre el conjunto de datos “datos_procesados”:

Ilustración 622

Estandarización de variables numéricas

```
# Estandarización con StandardScaler
scaler = StandardScaler()
datos_procesados[
    ['segundas_matriculas',
     'terceras_matriculas',
     'creditos']] = scaler.fit_transform(datos_procesados[
    ['segundas_matriculas',
     'terceras_matriculas',
     'creditos']])
```

Una vez se llevó a cabo la estandarización, se procede con la normalización de los mismos datos numéricos:

Ilustración 623

Normalización de variables numéricas

```
# Normalización
scaler = MinMaxScaler()
datos_procesados[
    ['segundas_matriculas',
     'terceras_matriculas',
     'creditos']] = scaler.fit_transform(datos_procesados[
    ['segundas_matriculas',
     'terceras_matriculas',
     'creditos']])
```

Una vez finalizada la normalización y estandarización, se tiene como resultado la siguiente vista previa del conjunto de datos procesado:

Ilustración 624

Vista previa de los datos estandarizados y normalizados

```
datos_procesados.head()
```

	etnia	nivel	segundas_matriculas	terceras_matriculas	facultad	docente	area_asignatura	carrera	aprobado	creditos
4	Mestizo	07	0.0	0.0	Facultad Economía	D41	3	Economía	S	0.15
5	Mestizo	07	0.0	0.0	Facultad Economía	D396	3	Economía	S	0.15
6	Mestizo	07	0.0	0.0	Facultad Economía	D776	3	Economía	S	0.15
7	Mestizo	07	0.0	0.0	Facultad Economía	D506	3	Economía	S	0.15
8	Mestizo	07	0.0	0.0	Facultad Economía	D629	3	Economía	S	0.15

En contraste, la copia del conjunto de datos antes de ser procesado tiene el siguiente contenido:

Ilustración 625

Vista previa de los datos sin estandarizar ni normalizar

```
temp.head()
```

	etnia	nivel	segundas_matriculas	terceras_matriculas	facultad	docente	area_asignatura	carrera	aprobado	creditos
4	Mestizo	07	0.0	0.0	Facultad Economía	D41	3	Economía	S	3.0
5	Mestizo	07	0.0	0.0	Facultad Economía	D396	3	Economía	S	3.0
6	Mestizo	07	0.0	0.0	Facultad Economía	D776	3	Economía	S	3.0
7	Mestizo	07	0.0	0.0	Facultad Economía	D506	3	Economía	S	3.0
8	Mestizo	07	0.0	0.0	Facultad Economía	D629	3	Economía	S	3.0

Por lo que, como se puede visualizar, los valores 3.0 correspondientes a la variable creditos hallados en los datos sin estandarizar ni normalizar, ahora son 0.15; esta misma dinámica aplica para los diferentes créditos existentes en los conjuntos de datos, estos, podrán ser observados a continuación con ayuda de un respectivo código:

Ilustración 626

Parte 1 de la visualización del contraparte de los datos estandarizados y normalizados de la variable credits

```
# Se itera por todos los valores únicos de la variable en el reporte estandarizado y normalizado
for credits in datos_procesados["credits"].unique():
    # Se muestra el valor estandarizado y normalizado
    print(f"Considernado el valor {round(credits,6)} del reporte estandarizado y normalizado.")
    print(f"Corresponde a: ", end="")
    # Se muestra el valor original
    print(temp[datos_procesados["credits"] == credits].iloc[0]["credits"], end="")
    print(" del reporte original.", end="\n\n")
```

Considernado el valor 0.15 del reporte estandarizado y normalizado.
Corresponde a: 3.0 del reporte original.

Considernado el valor 0.225 del reporte estandarizado y normalizado.
Corresponde a: 4.5 del reporte original.

Considernado el valor 0.2 del reporte estandarizado y normalizado.
Corresponde a: 4.0 del reporte original.

Considernado el valor 0.0 del reporte estandarizado y normalizado.
Corresponde a: 0.0 del reporte original.

Considernado el valor 0.3 del reporte estandarizado y normalizado.
Corresponde a: 6.0 del reporte original.

Considernado el valor 0.1 del reporte estandarizado y normalizado.
Corresponde a: 2.0 del reporte original.

Considernado el valor 0.375 del reporte estandarizado y normalizado.
Corresponde a: 7.5 del reporte original.

Considernado el valor 0.4 del reporte estandarizado y normalizado.
Corresponde a: 8.0 del reporte original.

Considernado el valor 0.25 del reporte estandarizado y normalizado.
Corresponde a: 5.0 del reporte original.

Considernado el valor 0.5 del reporte estandarizado y normalizado.
Corresponde a: 10.0 del reporte original.

Considernado el valor 0.35 del reporte estandarizado y normalizado.
Corresponde a: 7.0 del reporte original.

Considernado el valor 0.05 del reporte estandarizado y normalizado.
Corresponde a: 1.0 del reporte original.

Ilustración 627

Parte final de la visualización del contraparte de los datos estandarizados y normalizados de la variable credits

Considernado el valor 0.6 del reporte estandarizado y normalizado.
Corresponde a: 12.0 del reporte original.

Considernado el valor 0.625 del reporte estandarizado y normalizado.
Corresponde a: 12.5 del reporte original.

Considernado el valor 0.45 del reporte estandarizado y normalizado.
Corresponde a: 9.0 del reporte original.

Considernado el valor 0.55 del reporte estandarizado y normalizado.
Corresponde a: 11.0 del reporte original.

Considernado el valor 1.0 del reporte estandarizado y normalizado.
Corresponde a: 20.0 del reporte original.

Considernado el valor 0.95 del reporte estandarizado y normalizado.
Corresponde a: 19.0 del reporte original.

Considernado el valor 0.75 del reporte estandarizado y normalizado.
Corresponde a: 15.0 del reporte original.

Considernado el valor 0.1875 del reporte estandarizado y normalizado.
Corresponde a: 3.75 del reporte original.

Como se puede observar, cada crédito tiene un correspondiente valor único en el conjunto de datos normalizado y estandarizado, este mismo funcionamiento aplica para las demás variables numéricas, estas, de igual forma podrán visualizarse a continuación:

Ilustración 628

Visualización del contraparte de los datos estandarizados y normalizados de la variable `segundas_matriculas`

```
# Se itera por todos los valores únicos de la variable en el reporte estandarizado y normalizado
for segundas in datos_procesados["segundas_matriculas"].unique():
    # Se muestra el valor estandarizado y normalizado
    print(f"Considernado el valor {round(segundas,6)} del reporte estandarizado y normalizado.")
    print(f"Corresponde a: ", end="")
    # Se muestra el valor original
    print(temp[datos_procesados["segundas_matriculas"] == segundas].iloc[0]["segundas_matriculas"], end="")
    print(" del reporte original.", end="\n\n")
```

Considernado el valor 0.0 del reporte estandarizado y normalizado.
Corresponde a: 0.0 del reporte original.

Considernado el valor 0.166667 del reporte estandarizado y normalizado.
Corresponde a: 1.0 del reporte original.

Considernado el valor 0.333333 del reporte estandarizado y normalizado.
Corresponde a: 2.0 del reporte original.

Considernado el valor 0.5 del reporte estandarizado y normalizado.
Corresponde a: 3.0 del reporte original.

Considernado el valor 0.833333 del reporte estandarizado y normalizado.
Corresponde a: 5.0 del reporte original.

Considernado el valor 0.666667 del reporte estandarizado y normalizado.
Corresponde a: 4.0 del reporte original.

Considernado el valor 1.0 del reporte estandarizado y normalizado.
Corresponde a: 6.0 del reporte original.

Ilustración 629

Visualización del contraparte de los datos estandarizados y normalizados de la variable `terceras_matriculas`

```
# Se itera por todos los valores únicos de la variable en el reporte estandarizado y normalizado
for terceras in datos_procesados["terceras_matriculas"].unique():
    # Se muestra el valor estandarizado y normalizado
    print(f"Considernado el valor {round(terceras,6)} del reporte estandarizado y normalizado.")
    print(f"Corresponde a: ", end="")
    # Se muestra el valor original
    print(temp[datos_procesados["terceras_matriculas"] == terceras].iloc[0]["terceras_matriculas"], end="")
    print(" del reporte original.", end="\n\n")
```

Considernado el valor 0.0 del reporte estandarizado y normalizado.
Corresponde a: 0.0 del reporte original.

Considernado el valor 0.2 del reporte estandarizado y normalizado.
Corresponde a: 1.0 del reporte original.

Considernado el valor 0.4 del reporte estandarizado y normalizado.
Corresponde a: 2.0 del reporte original.

Considernado el valor 0.6 del reporte estandarizado y normalizado.
Corresponde a: 3.0 del reporte original.

Considernado el valor 1.0 del reporte estandarizado y normalizado.
Corresponde a: 5.0 del reporte original.

Una vez analizado y comprendido los valores que corresponden a las variables numéricas de los datos que fueron utilizados para la creación del modelo, ha llegado el momento de crear

diccionarios donde se recopilen los datos originales con sus respectivas estandarizaciones y normalizaciones para su posterior exportación, esto, con la finalidad de que cuando se ingresen estos datos del estudiante en el aplicativo web, se codifiquen a los datos hallados en el modelo, y, adicionalmente, permitir al usuario mayor facilidad en el ingreso de datos.

Ilustración 630

Creación de diccionarios con los valores únicos de las variables numéricas y sus valores estandarizados y normalizados

```
# Se define una lista que almacenará todos los diccionarios
diccionarios_para_exportar = []

# Se itera por todas las variables numéricas
for variable in datos_procesados.describe().columns:
    # Se define un diccionario temporal
    dict_temp = {}

    # Se itera por todos los valores únicos de la variable en el reporte estandarizado y normalizado
    for dato in datos_procesados[variable].unique():
        # Se toma el valor estandarizado y normalizado
        procesado = dato
        # Se toma el valor original correspondiente al estandarizado y normalizado
        original = temp[datos_procesados[variable] == dato].iloc[0][variable]

        # Se agrega el par de valores en el diccionario temporal
        dict_temp[original] = procesado

    # Se agrega el diccionario a la lista de diccionarios
    diccionarios_para_exportar.append(dict_temp)
    # Se elimina el diccionario temporal
    del dict_temp

# Se comprueba que se ha ingresado correctamente la información
print(f"Existen {len(diccionarios_para_exportar)} diccionarios.")
print(f"Existen {len(datos_procesados.describe().columns)} variables numéricas")

Existen 3 diccionarios.
Existen 3 variables numéricas
```

Como se puede observar, se ha logrado guardar exitosamente los posibles valores de las distintas categorías de las variables numéricas, esto significa que en este momento se puede realizar este proceso sobre las variables categóricas. Para empezar, es necesario comentar que no se realizará ningún proceso de exportación sobre la variable “aprobado” debido a que esta no se ingresará por el usuario, es decir, esta variable es el resultado de la predicción del modelo, sin embargo, es necesario codificarla para que se permita la codificación de la variable docente.

Ilustración 631

Codificación de la variable aprobado

```
datos_procesados['aprobado'] = datos_procesados['aprobado'].map({'S': 1, 'N': 0})
```

Por otra parte, el proceso de codificación que se llevó a cabo sobre las variables “etnia”, “nivel”, “facultad”, “area_asignatura” y “carrera” fue empleado usando la técnica de One-Hot-Encoding, la misma fue explicada en su momento, sin embargo, de manera resumida, la técnica crea campos por cada categoría de las halladas en los datos de cada variable, donde coloca 1 si esta fue la que perteneció al registro y 0 a las demás; es decir, se crean columnas por cada facultad que existe, pero, si el estudiante pertenece a una facultad A, exclusivamente la nueva columna A tendrá un valor de 1, mientras que las demás poseerán un valor de 0. Esto significa que el proceso de exportación de estas variables en específico es enfocado únicamente en la exportación de las categorías existentes en ellas, por lo que, antes de proceder a ejecutar la técnica de One-Hot-Encoding, es necesario realizar recopilar las categorías existentes en cada variable mencionada en esta sección:

Ilustración 632

Creación de una lista con las categorías de las variables categóricas con excepción de la variable docente

```
# Se toman todas las variables categóricas
categoricas = datos_procesados.describe(include=object).columns

# Se crea una lista que almacenará todas las categorías de las variables
categorias = []

# Se itera sobre las variables categóricas
for variable in categoricas:
    # Se evita interactuar con la variable docente por el momento
    if variable == "docente":
        continue

    # Se recopilan todas las categorías de la variable considerada
    categorias_temp = datos_procesados[variable].unique().tolist()

    # Se añade la lista de categorías de la variable considerada
    # a la lista de todas las categorías de todas las variables
    categorias.append(categorias_temp)

# Se comprueba que se hayan ingresado correctamente los datos
print(f"Existen {len(categorias)} listados de categorías", end=" ")
print(f"y {len(categoricas)} variables categóricas.")

count = 0
for variable in categoricas:
    # Se evita el contacto con la variable docente por el momento
    if variable == "docente":
        continue
    # Se imprime la comprobación.
    print(f"Existen {len(datos_procesados[variable].unique())} categorías en la variable {variable},", end=" ")
    print(f"lo que se asemeja a las {len(categorias[count])} categoricas guardadas.")
    count += 1
```

Existen 5 listados de categorías y 6 variables categóricas.
Existen 8 categorías en la variable etnia, lo que se asemeja a las 8 categoricas guardadas.
Existen 10 categorías en la variable nivel, lo que se asemeja a las 10 categoricas guardadas.
Existen 14 categorías en la variable facultad, lo que se asemeja a las 14 categoricas guardadas.
Existen 10 categorías en la variable area_asignatura, lo que se asemeja a las 10 categoricas guardadas.
Existen 71 categorías en la variable carrera, lo que se asemeja a las 71 categoricas guardadas.

Una vez recopilado las diferentes categorías de las variables categóricas con excepción de la variable “docente” y haber comprobado que fue realizado con éxito el procedimiento, se puede llevar a cabo la aplicación de la técnica One-Hot-Encoding:

Ilustración 633

Ejecución de la técnica de One-Hot-Encoding

```
# Se aplica la técnica de one-hot-encoding
datos_procesados = pd.get_dummies(datos_procesados,
                                   columns=["nivel",
                                           "area_asignatura",
                                           'etnia',
                                           'facultad',
                                           "carrera"])
```

Una vez se ha llevado a cabo la codificación de la gran mayoría de variables categóricas, es necesario llevar a cabo la codificación de la variable docente, para esta ocasión, como se mencionó anteriormente, se llevó a cabo mediante la técnica de “target_encoding”, el funcionamiento de esta consiste en agrupar los datos por categorías, calcular la media con respecto a la variable objetivo, es decir, “aprobado”, y sustituir cada categoría por la media calculada; esto quiere decir, que para exportar las categorías de la variable “docente” es necesario realizarlo mediante pares, es decir, diccionarios de manera similar a como se realizó para las variables numéricas. Sin embargo, es necesario comentar que en la ilustraciones siguiente donde se observa la identificación de cada docente dado su codificación, exclusivamente se mostrará una breve parte del resultado, esto, debido a que existe una gran cantidad de docentes.

Ilustración 634

Visualización del contraparte de los datos codificados de la variable docente

```
for docente in datos_procesados["docente"].unique():
    # Se muestra el valor estandarizado y normalizado
    print(f"Considernado el valor {docente} del reporte estandarizado y normalizado.")
    print(f"Corresponde a: ", end="")
    # Se muestra el valor original
    print(temp2[datos_procesados["docente"] == docente].iloc[0]["docente"], end="")
    print(" del reporte original.", end="\n\n")
```

Considernado el valor 0.790254585641167 del reporte estandarizado y normalizado.
Corresponde a: D41 del reporte original.

Considernado el valor 0.7694270019106654 del reporte estandarizado y normalizado.
Corresponde a: D396 del reporte original.

Considernado el valor 0.6238457306186898 del reporte estandarizado y normalizado.
Corresponde a: D776 del reporte original.

Considernado el valor 0.8482128741630813 del reporte estandarizado y normalizado.
Corresponde a: D506 del reporte original.

Considernado el valor 0.6938671226631044 del reporte estandarizado y normalizado.
Corresponde a: D629 del reporte original.

Considernado el valor 0.6245150407872241 del reporte estandarizado y normalizado.
Corresponde a: D642 del reporte original.

Considernado el valor 0.8553828667186221 del reporte estandarizado y normalizado.
Corresponde a: D512 del reporte original.

Considernado el valor 0.8906238498498837 del reporte estandarizado y normalizado.
Corresponde a: D522 del reporte original.

Dado lo observable, se puede asegurar que se puede llevar a cabo un procedimiento similar al realizado sobre las variables numéricas con el mismo diccionario, el código y su resultado se puede apreciar a continuación:

Ilustración 635

Agregación de los valores únicos de la variable docente y sus valores codificados al diccionario “diccionario_para_exportar”

```
# Se define un diccionario temporal
dict_temp = {}

# Se itera por todos los valores únicos de la variable en el reporte estandarizado y normalizado
for dato in datos_procesados["docente"].unique():
    # Se toma el valor estandarizado y normalizado
    procesado = dato
    # Se toma el valor original correspondiente al estandarizado y normalizado
    original = temp[datos_procesados["docente"] == dato].iloc[0]["docente"]

    # Se agrega el par de valores en el diccionario temporal
    dict_temp[original] = procesado

# Se agrega el diccionario a la lista de diccionarios por exportar
diccionarios_para_exportar.append(dict_temp)

# Se comprueba que se ha ingresado correctamente la información
print(f"Existen {len(diccionarios_para_exportar[-1])} datos en el diccionario de docentes.")
print("Existen", len(datos_procesados["docente"].unique()), "diferentes docentes en los reportes oficiales.")

Existen 596 datos en el diccionario de docentes.
Existen 596 diferentes docentes en los reportes oficiales.
```

Finalmente, ha llegado el momento de exportar todas las categorías y codificaciones que se han ido guardando en el diccionario “diccionarios_por_exportar” y la lista “categorias”, para ello, se creará un documento “.txt” con todos los datos recopilados, el código puede apreciarse a continuación junto a una vista previa del documento creado:

Ilustración 636

Código para la creación de un documento ".txt" y el guardado de todas las categorías

```
# Se crea un documento llamado "datos_exportados.txt"
with open("datos_exportados.txt", "w") as file:
    # Se escriben los datos hallados en el diccionario
    for diccionario in diccionarios_para_exportar:
        for items in diccionario.items():
            file.write(str(items) + "\n")
        # Se identifica el final de una categoría por "end"
        file.write("end\n")

    # Se escriben los datos de las variables categóricas
    for categoria in categorias:
        for item in categoria:
            file.write(str(item) + "\n")
        # Se identifica el final de una categoría por "end"
        file.write("end\n")
```


Ilustración 637

Vista previa del documento creado

 datos_exportados: Bloc de notas

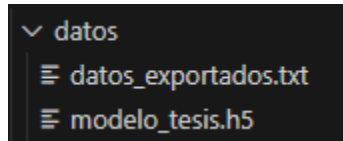
Archivo	Edición	Formato	Ver	Ayuda
---------	---------	---------	-----	-------

```
(0.0, 0.0)
(1.0, 0.16666666666666669)
(2.0, 0.3333333333333334)
(3.0, 0.5)
(5.0, 0.8333333333333334)
(4.0, 0.6666666666666667)
(6.0, 1.0)
end
(0.0, 0.0)
(1.0, 0.2)
(2.0, 0.39999999999999997)
(3.0, 0.6)
(5.0, 1.0)
end
(3.0, 0.15000000000000002)
(4.5, 0.22500000000000003)
(4.0, 0.20000000000000004)
(0.0, 0.0)
(6.0, 0.30000000000000004)
(2.0, 0.1)
(7.5, 0.37500000000000006)
(8.0, 0.40000000000000001)
(5.0, 0.25000000000000006)
(10.0, 0.5000000000000001)
(7.0, 0.35000000000000001)
(1.0, 0.05)
(12.0, 0.6000000000000001)
(12.5, 0.6250000000000001)
(9.0, 0.45000000000000007)
(11.0, 0.5500000000000002)
(20.0, 1.0000000000000002)
(19.0, 0.9500000000000001)
(15.0, 0.75)
(3.75, 0.18750000000000003)
end
('D41', 0.790254585641167)
('D396', 0.7694270019106654)
('D776', 0.6238457306186898)
('D506', 0.8482128741630813)
('D629', 0.6938671226631044)
```


Una vez se realizó la exportación de todas las categorías de las distintas variables y del modelo, es necesario mover los archivos resultantes a la carpeta donde se programará el aplicativo web, este proceso puede observar a continuación:

Ilustración 638

Vista de la ubicación de los archivos donde se realizará el aplicativo web



Una vez se ha logrado copiar los archivos requeridos, es decir, las categorías de cada variable y el modelo de machine learning de la cuarta propuesta de red neuronal, es momento de comenzar la programación del aplicativo web. Para esto, primero se creará un módulo llamado “trados.py”, el mismo ha sido desarrollado por el investigador y poseerá una clase con el mismo nombre, es decir, tradatos; el módulo posee varias funciones que serán descritas a continuación:

Ilustración 639

Descripción del modulo e importación de "pandas"

```
1  '''
2
3  Modulo de tratamiento de datos
4
5  Desarrollado por: Sebastián Felipe Tamayo Proaño
6
7  Descripción: El modulo contiene una clase con métodos capaces de:
8      - importar datos
9      - retornar categorías codificadas
10     - retornar categorías sin codificar
11     - retornar categorías
12     - preparar datos según un valor dado
13     - preparar datos ingresados para la propuesta de la predicción
14
15  '''
16
17  import pandas as pd
18
```

Para empezar, es visible la descripción del módulo junto al autor de este, además, para este módulo se empleó la librería de “pandas”.

Ilustración 640

Definición de la clase, el constructor y la función importar_datos

```
19 class tradatos():
20
21     def __init__(self):
22         # Se importan los datos de las categorías
23         self.importar_datos()
24         # Se crea un diccionario que poseerá todas las variables y sus categorías
25         self.variables_dict = {}
26         # Se clasifican los datos ingresados
27         self.clasificar_datos()
28
29     def importar_datos(self):
30         # Se importan los datos exportados de la fase de modelado
31         with open("datos\datos_exportados.txt", "r") as file:
32             self.datos = file.readlines()
33
```

Este módulo consiste en una sola clase llamada “tradatos”, la misma viene de su sentido, el tratamiento de los datos, en la ilustración es visible la definición del constructor de la clase en donde se emplea la función importar_datos cuya definición está ubicada posterior al constructor, lo que realiza este método es importar los datos del documento “datos_exportados.txt” en la variable datos, por otra parte, en el constructor se crea un diccionario que almacenará todos las variables y sus categorías, finalmente, se llama al método clasificar_datos, esta será explicada a continuación:

Ilustración 641

Definición de la función clasificar_datos

```
34 def clasificar_datos(self):
35     '''
36     Clasifica los datos por categoría en el diccionario variables_dict
37     '''
38     # Se crea una lista con las variables en el orden en el que fueron exportadas
39     variables = ["segundas_matriculas", "terceras_matriculas",
40                 "creditos", "docente", "etnia",
41                 "nivel", "facultad", "area_asignatura",
42                 "carrera"]
43     # Se identifica el orden con el que aparecerán las variables
44     orden = 0
45     # Almacena las categorías existentes en cada variable por vez
46     categorias = []
47     # Se clasifican las categorías por variable
48     for categoria in self.datos:
49         # Se elimina el caracter "\n"
50         categoria = categoria.strip()
51         # Se evalúa el cambio de categoría
52         if categoria == "end":
53             # Se crea un nuevo ítem en el diccionario "variables" con la categoría correspondiente
54             self.variables_dict[variables[orden]] = categorias
55             # Se resetea el contenido de categorías
56             categorias = []
57             # Se cambia de variable
58             orden += 1
59         else:
60             # Se formatea el contenido en caso de estar separado por comas y en paréntesis
61             if "(" in categoria:
62                 temp = categoria[1:-1].split(",")
63                 temp[0] = temp[0].strip().strip("'")
64                 temp[1] = temp[1].strip().strip("'")
65                 # Se agrega a la lista para después agregarse al diccionario
66                 categorias.append(temp)
67             else:
68                 # Se agrega a la lista para después agregarse al diccionario
69                 categorias.append(categoria)
```

El método `clasificar_datos` tiene el propósito de clasificar los datos importados del documento `"datos_exportados.txt"`, lo que realiza es evaluar cada uno de los datos hallados en la variable `"datos"` y ubica las categorías dado el orden en el que se fue ingresada la información en el documento `".txt"` hasta encontrarse con `"end"`, al hacerlo, se continúa con la siguiente variable, y así sucesivamente hasta terminar con todas las variables. Cabe recalcar que el orden en el que se guardaron los datos es el hallado en la lista `"variables"`.

Ilustración 642

Definición de las funciones get_sincodificar, get_codificadas y get_datos

```
71 def get_sincodificar(self, variable):
72     """
73     Esta función devuelve los valores sin codificar de las variables
74     que fueron sujetas a codificaciones, es decir, las variables numéricas y
75     la variable docente.
76     """
77     if variable in ["segundas_matriculas", "terceras_matriculas", "creditos", "docente"]:
78         sincodificar = []
79         for dato in self.variables_dict[variable]:
80             sincodificar.append(dato[0])
81         return sincodificar
82     return f"Este método no aplica a la variable {variable}."
83
84 def get_codificadas(self, variable, item):
85     """
86     Esta función devuelve el valor codificado de un item ingresado
87     dada la variable
88     """
89     if variable in ["segundas_matriculas", "terceras_matriculas", "creditos", "docente"]:
90         codificadas = []
91         for dato in self.variables_dict[variable]:
92             if dato[0] == item:
93                 return dato[1]
94     return f"No tiene codificación este valor: {item} de la variable {variable}."
95
96 def get_datos(self, variable):
97     """
98     Esta función devuelve las categorías de una variable dada
99     """
100     return self.variables_dict[variable]
101
```

La función get_sincodificar, devuelve la versión no codificada de las variables numéricas y de la variable “docente”, mientras que, el método get_codificadas realiza lo opuesto, devuelve la información codificada de las variables numéricas y de la variable “docente”. Por otra parte, el método get_datos funciona para retornar los datos de las demás variables categóricas.

Ilustración 643

Definición de la función get_seleccionado

```
100 def get_seleccionado(self, variable, item):
101     """
102     Esta función devuelve una lista con valores asignados en las categorías
103     de las variables categóricas con excepción de la variable docente.
104     El item ingresado será el único valor con un 1 mientras que las demás categorías
105     poseerán un valor de 0.
106     """
107     if variable not in ["segundas_matriculas", "terceras_matriculas", "creditos", "docente"]:
108         temp_list = []
109         for dato in self.variables_dict[variable]:
110             if dato == item:
111                 temp_list.append([dato,1])
112             else:
113                 temp_list.append([dato,0])
114         return temp_list
115     else:
116         return f"Esta variable es admisible para este proceso."
117
```

Este método es muy importante debido a que simula el funcionamiento de la técnica One-Hot-Encoding, está dada la selección de una opción de una de las categorías de una las variables categóricas con excepción de la variable “docente”, crea una lista de pares donde coloca el nombre de la categoría al lado de un valor, un valor de 1 si es la categoría seleccionada mientras que en las demás categorías de la variable se coloca un valor de 0, esta lista luego es retornada.

Ilustración 644

Primera parte de la definición de la función crear_propuesta_prediccion

```
118 def crear_propuesta_prediccion(self, datos):
119     """
120     Esta función tiene como objetivo crear un dataframe con los datos necesarios para realizar
121     una predicción.
122     Acepta una lista con los datos ingresados para cada variable.
123     """
124     # Se crea una lista con las variables en el orden en el que fueron exportadas
125     variables = ["segundas_matriculas", "terceras_matriculas",
126                 "creditos", "docente", "etnia",
127                 "nivel", "facultad", "area_asignatura",
128                 "carrera"]
129
130     # Se crea un diccionario temporal que se usará para la creación del dataframe
131     dict_for_pred = {}
132
133     # Se organiza la información
134     for i in range(len(datos)):
135         # Se organiza la información para las variables numéricas y docente
136         if i < 4:
137             valor = self.get_codificadas(variables[i], datos[i])
138             dict_for_pred[variables[i]] = valor
139         # Se organiza la información para las variables categóricas restantes
140         else:
141             valores = self.get_seleccionado(variables[i], datos[i])
142             for valor in valores:
143                 # Se ajusta el nombre del item en el diccionario
144                 nombre_ajustado = variables[i] + "_" + valor[0]
145                 dict_for_pred[nombre_ajustado] = valor[1]
146
147     # Se crea una Serie con base al diccionario creado al comienzo
148     df = pd.DataFrame(dict_for_pred, index=[0])
149
```

Esta es la primera parte del método más importante del módulo, en específico, esta primera parte se encarga de organizar la información para la generación de la predicción, en un comienzo se crean todas las columnas necesarias para completar las 117 variables requeridas que fueron utilizadas para la creación del modelo, como se puede observar, en muchas ocasiones se ajustan los nombres según los requerimientos para que se concreten según se encuentran en el conjunto de datos de entrenamiento, por ejemplo, las variables relacionadas a etnia tienen un formato de esta forma:

Ilustración 645

Vista el formato de nombres de la variable etnia

etnia_Afroecuatoriano	etnia_Blanco	etnia_Indígena	etnia_Mestizo	etnia_Montubio	etnia_Mulato
0	0	0	1	0	0

Ilustración 646

Parte final de la definición de la función crear_propuesta_prediccion

```
150 # Se arregla el orden de la serie conforme a como fue en la creación del modelo
151 nuevo_orden = ["segundas_matriculas", "terceras_matriculas", "docente", "creditos"]
152 # Se organizan los demás nombres que fueron asignados dado el contenido de la categoría
153 # Para nivel
154 nuevo_orden.extend(["nivel_0" + str(i) if i < 10 else "nivel_" + str(i) for i in range(1,11)])
155 # Para area_asignatura
156 nuevo_orden.extend(["area_asignatura_" + str(i) for i in range(1,12) if i != 8])
157 # Para etnia
158 etnias = self.get_datos("etnia")
159 etnias.sort()
160 nuevo_orden.extend(["etnia_" + etnia for etnia in etnias])
161 # Para facultad
162 facultades = self.get_datos("facultad")
163 facultades.sort()
164 nuevo_orden.extend(["facultad_" + facultad for facultad in facultades])
165 # Para carreras
166 carreras = self.get_datos("carrera")
167 carreras.sort()
168 nuevo_orden.extend(["carrera_" + carrera for carrera in carreras])
169
170 # Se cambia el orden de las variables en el dataframe
171 df = df.reindex(columns=nuevo_orden)
172
173 # Se cambia el tipo de datos de las variables
174 df[["segundas_matriculas", "terceras_matriculas", "docente", "creditos"]] = df[["segundas_matriculas",
175 "terceras_matriculas",
176 "docente",
177 "creditos"]].astype(float)
178
179 # Se retorna el dataframe creado para la predicción
180 return df
```

Una vez visualizada la primera parte del método crear_propuesta_prediccion, es momento de comentar con respecto a la parte final, y, para comenzar se debe mencionar lo crucial que resulta cada línea de código, en un comienzo el objetivo de esta sección es reordenar las columnas para que se ajusten al orden con el que se entrenó el modelo de predicción de rendimiento académico, por lo que, de la línea 152 hasta la línea 170, se recopilan los nombres de todas las variables en el orden correcto, posterior a ello, en la línea 173 se ejecuta el reordenamiento de las columnas, y, finalmente, se ajusta el tipo de datos para que se corresponda

con los tipos de datos que fueron usados para la creación del modelo; posterior a ello se procede a retornar el dataframe resultante.

En este punto, es momento de programar el aplicativo web, en esta ocasión se ha decidido realizar el despliegue del modelo de predicción de rendimiento académico utilizando Flask que consiste en un micro marco de trabajo de desarrollo web, el cual, se consideró propicio para el objetivo de desplegar el modelo realizado. El desarrollo será comentado junto a sus ilustraciones a continuación:

Ilustración 647

Importación de módulos, definiciones iniciales y declaración del modelo de la página principal

```
1  from flask import Flask, render_template, request
2  from keras.models import load_model
3  import modulos.tradatos as tradatos
4
5  app = Flask(__name__)
6
7  # Se carga los datos
8  datos = tradatos.tradatos()
9
10 # Se carga el modelo
11 modelo = load_model("datos\modelo_tesis.h5")
12
13 # Página principal
14 @app.route("/")
15 @app.route("/index")
16 def home():
17     return render_template("index.html")
18
```

Para empezar, es necesario comentar que Flask funciona mediante vistas, plantillas y solicitudes, en esta ocasión, se definirán las vistas o modelos hallados en el documento “sprast.py”, que, es de interés mencionar que el nombre del documento no es casualidad y será explicado en los resultados del despliegue a mayor detalle. Por lo demás, en el código es visible en primer lugar la importación Flask, render_template y request, estos módulos de la librería flask permitirán la creación del aplicativo web, la navegación y el envío de solicitudes del

usuario como del servidor; por otra parte, el módulo “load_model” de la librería keras.models, permitirá la importación del modelo de la quinta propuesta de red neuronal, finalmente, es visible que se importa el módulo creado para el tratamiento de datos.

Por otro lado, el código en la anterior ilustración permite la instanciación de la aplicación web, la carga de los datos, es decir, la inicialización de la clase tradatos; la importación del modelo y la definición de la primera ruta o modelo del aplicativo/página web.

Ilustración 648

Parte 1 de la definición del modelo de la página "predicción"

```
19 @app.route("/prediccion", methods=["GET", "POST"])
20 def prediccion():
21     if request.method == "POST":
22         # Se recopila la información ingresada
23         nombre = request.form['nombre']
24         etnia = request.form['etnia']
25         nivel = request.form['semestre']
26         facultad = request.form['facultad']
27         carrera = request.form['carrera']
28         segunda_matricula = request.form['segunda_matricula']
29         tercera_matricula = request.form['tercera_matricula']
30         asignatura = request.form['asignatura']
31         area_asignatura = request.form['area_asignatura']
32         credito = request.form['credito']
33         docente = request.form['docente']
34
35         # Se organiza la información en el orden requerido
36         datos_recopilados = [segunda_matricula, tercera_matricula, credito,
37                               docente, etnia, nivel, facultad,
38                               area_asignatura, carrera]
39
40         # Se genera el dataframe para la predicción
41         X_test = datos.crear_propuesta_prediccion(datos_recopilados)
42
43         # Se realiza la predicción con el modelo de la quinta propuesta de red neuronal
44         y_pred_prob = modelo.predict(X_test)
45         y_pred = (y_pred_prob > 0.5).astype(int)
46
47         # Se mejora la presentación de resultado
48         if y_pred[0][0] == 1:
49             resultado = "aprobar"
50         else:
51             resultado = "reprobar"
52
53         # Se muestra el resultado en una nueva página
54         return render_template('resultado.html', resultado=resultado, nombre=nombre, asignatura=asignatura)
```

En diferente sentido, el código visible en la presente ilustración corresponde a la primera parte del modelo o ruta de la página “prediccion.html”, en esta página le permitirá al usuario ingresar datos del estudiante quien será sujeto a una predicción, por lo que, en esta primera parte

del código, se puede observar que acepta una solicitud de “POST”, se recopila la información ingresada por el usuario, se organiza la información para llamar al método “crear_propuesta_prediccion” y, finalmente, se genera la predicción del estudiante ingresado.

Ilustración 649

Parte final de la definición del modelo de la página "predicción"

```
56 # Se cargan las opciones a elegir de las diferentes categorías
57 etnias = datos.get_datos("etnia")
58 semestres = datos.get_datos("nivel")
59 facultades = datos.get_datos("facultad")
60 carreras = datos.get_datos("carrera")
61 segundas_matriculas = datos.get_sincodificar("segundas_matriculas")
62 terceras_matriculas = datos.get_sincodificar("terceras_matriculas")
63 areas_asignaturas = datos.get_datos("area_asignatura")
64 creditos = datos.get_sincodificar("creditos")
65 docentes = datos.get_sincodificar("docente")
66
67 # Se modifica el formato de algunas listas para mejorar la visibilidad
68 areas_asignaturas = [int(x) for x in areas_asignaturas]
69 creditos = [float(x) for x in creditos]
70
71 # Se ordenan los datos para una mejor visibilidad
72 etnias.sort()
73 semestres.sort()
74 facultades.sort()
75 carreras.sort()
76 segundas_matriculas.sort()
77 terceras_matriculas.sort()
78 areas_asignaturas.sort()
79 creditos.sort()
80 creditos.sort()
81 docentes.sort()
82
83 # Se carga la página
84 return render_template("prediccion.html",
85                        etnias=etnias, semestres=semestres, facultades=facultades,
86                        carreras=carreras, segundas_matriculas=segundas_matriculas,
87                        terceras_matriculas=terceras_matriculas, areas_asignaturas=areas_asignaturas,
88                        creditos=creditos, docentes=docentes)
```

En contraste, la segunda parte del código del módulo de la página de ingreso de los datos del estudiante para su predicción consiste en permitir que el usuario pueda seleccionar las opciones de cada variable que fueron utilizadas para la creación del modelo, haciendo que incluso sea más fácil la interacción con el llenado de los datos.

Ilustración 650

Definición del modelo de la página "resultado"

```
90 # Página que permite visualizar el resultado de la predicción
91 @app.route("/resultado")
92 def resultado():
93     resultado = request.args.get("resultado", None)
94     nombre = request.args.get("nombre", None)
95     asignatura = request.args.get("asignatura", None)
96     return render_template("resultado.html", resultado=resultado, nombre=nombre, asignatura=asignatura)
97
98 if __name__ == "__main__":
99     app.run(debug=True)
```

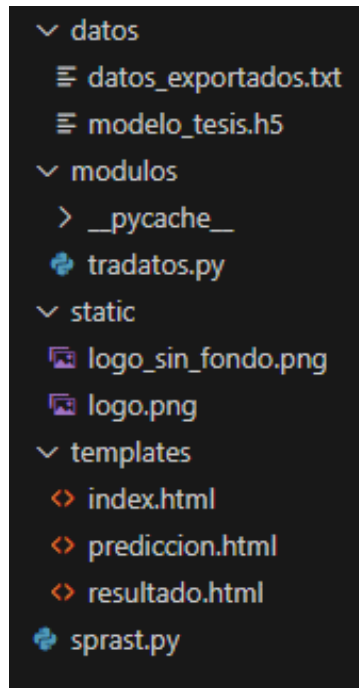
En último término, se puede visualizar que se define el último modelo que corresponde la página donde se podrá visualizar los resultados de la predicción, y, a partir de la línea 98, se procede a ejecutar el aplicativo web.

En este punto, se ha logrado la codificación de la lógica del aplicativo web, por lo que, se llevará a cabo la programación de las plantillas, es decir, de la interfaz de usuario; sin embargo, el código empleado para la interfaz no se visualizará mediante ilustraciones en este trabajo de titulación, no obstante, si el distinguido lector desea conocer a fondo el desarrollo de la interfaz de usuario, puede saciar su curiosidad en la sección de anexos, allí no solo se podrá encontrar con las tres diferentes páginas del aplicativo web, sino que también podrá tener acceso al modelo de la quinta propuesta de red neuronal.

De todas maneras, la carpeta donde se realizó el aplicativo web se encuentra constituido por lo siguiente:

Ilustración 651

Contenido del aplicativo web



Como se puede observar, un punto que no se conversó a lo largo de esta sección es la carpeta “static”, sin embargo, se obvió su detalle debido a que esta únicamente contiene los logos que fueron realizados para el despliegue del modelo.

7.2. Resultados y evaluación del despliegue

Para empezar, se ha decidió llamar al modelo de predicción de rendimiento académico “SPRAST”, cuyo nombre significa “Sistema de Predicción de Rendimiento Académico realizado por Sebastián Tamayo”, asimismo, el logo que se realizó para el modelo fue el siguiente:

Ilustración 652
Logo de SPRAST



Por otra parte, los resultado del despliegue del modelo fueron los siguientes:

Ilustración 653

Primera parte de la página de inicio de SPRAST

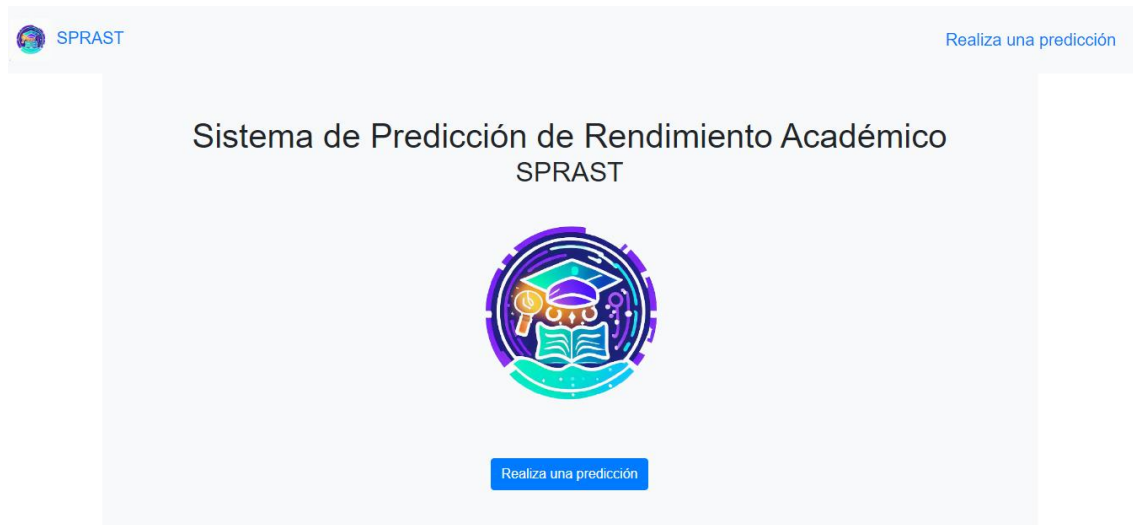


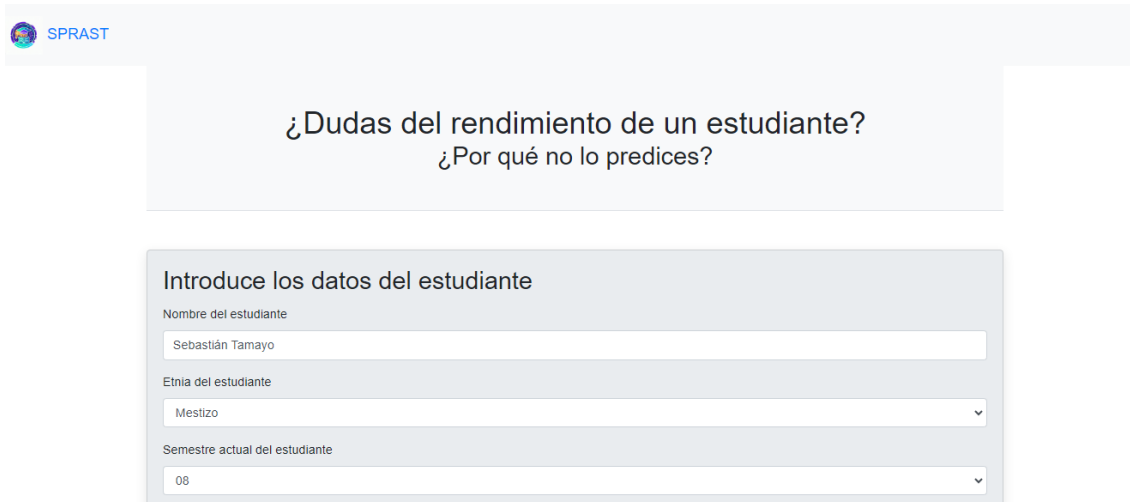
Ilustración 654

Parte final de la página de inicio de SPRAST



Ilustración 655

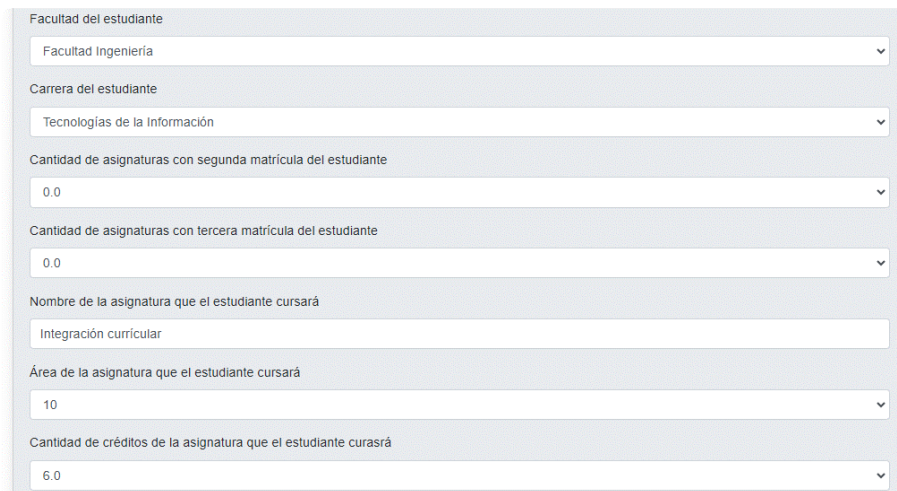
Parte 1 de la página de ingreso de datos del estudiante para su predicción



The screenshot shows the top section of a web form. At the top left is the SPRAST logo. In the center, there is a light gray box with the text: "¿Dudas del rendimiento de un estudiante?" and "¿Por qué no lo predices?". Below this is a section titled "Introduce los datos del estudiante". It contains three input fields: "Nombre del estudiante" with the value "Sebastián Tamayo", "Etnia del estudiante" with a dropdown menu showing "Mestizo", and "Semestre actual del estudiante" with a dropdown menu showing "08".

Ilustración 656

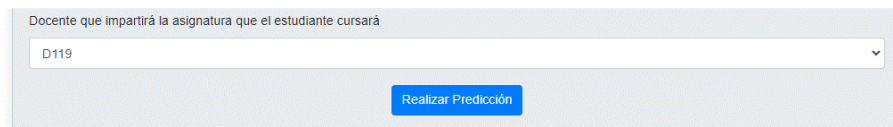
Parte 2 de la página de ingreso de datos del estudiante para su predicción



The screenshot shows the second section of the web form. It contains seven input fields: "Facultad del estudiante" (dropdown menu showing "Facultad Ingeniería"), "Carrera del estudiante" (dropdown menu showing "Tecnologías de la Información"), "Cantidad de asignaturas con segunda matrícula del estudiante" (dropdown menu showing "0.0"), "Cantidad de asignaturas con tercera matrícula del estudiante" (dropdown menu showing "0.0"), "Nombre de la asignatura que el estudiante cursará" (text input field showing "Integración curricular"), "Área de la asignatura que el estudiante cursará" (dropdown menu showing "10"), and "Cantidad de créditos de la asignatura que el estudiante cursará" (dropdown menu showing "6.0").

Ilustración 657

Parte final de la página de ingreso de datos del estudiante para su predicción



The screenshot shows the final section of the web form. It contains one input field: "Docente que impartirá la asignatura que el estudiante cursará" (dropdown menu showing "D119"). Below this field is a blue button labeled "Realizar Predicción".

Ilustración 658

Página donde se puede apreciar el resultado de la predicción

 SPRAST

Realiza una predicción

Resultados de la Predicción

Resultado de la Predicción

El estudiante **Sebastián Tamayo** tiene una alta probabilidad de **aprobar** la asignatura **Integración curricular**.

Realizar otra predicción

© 2023 Sebastián Tamayo. All rights reserved.

CAPÍTULO VIII: CONCLUSIONES

8. Conclusiones

8.1. Conclusión general

En general, se logró crear un modelo de machine learning capaz de predecir el rendimiento académico universitario de los estudiantes de la Pontificia Universidad Católica del Ecuador.

8.2. Conclusiones específicas

- Se concluyó que las variables referidas a la facultad a la que pertenece un estudiante, su carrera, su nivel académico, el área de asignatura que cursará o está considerando cursar, la cantidad de créditos que equivale la asignatura, el docente que imparte dicha asignatura, la etnia, su cantidad de asignaturas con segunda y tercera matrícula influyen directamente en el rendimiento académico de los estudiantes universitarios en la Pontificia Universidad Católica del Ecuador.
- Se lograron crear varios modelos de predicción de rendimiento académico con un excelso desempeño en su mayoría, es decir, dado el enfoque y gustos del investigador, puede tomarse el lujo de seleccionar cualquiera de los modelos realizados pues cumple con su labor bajo un criterio de éxito real.
- Se concluyó una posible relación entre el decrecimiento del rendimiento académico y el decrecimiento de los estudiantes matrículas en nuevos periodos académicos en la Pontificia Universidad Católica del Ecuador, e incluso con la educación durante la época en pandemia.
- Se concluyó que, la inexistencia de una relación entre el rendimiento académico y ser beneficiario de una beca socioeconómica podría deberse exclusivamente a los datos

utilizados para esta investigación, es decir, es probable que dado las situaciones culturales, históricas y geopolíticas que afectaron a la Pontificia Universidad Católica del Ecuador durante la época de la pandemia, repercutió en que la beca socioeconómica así como otros factores, hayan sido identificados como relaciones o no, que influyen en el rendimiento académico universitario.

8.3. Recomendaciones

- Se recomienda al investigador o lector de este trabajo de titulación, experimentar por su cuenta con diferentes enfoques e hipótesis, pues, en muchas ocasiones las investigaciones se ven sesgadas por sus investigadores, por lo que es muy probable que nuevas investigaciones fueren resultados diferentes, de todas maneras, se recomienda al uso de los datos que se han compartido como anexo, el mismo se llama “datos_para_el_investigador.csv”.
- Se recomienda el utilizar una mayor cantidad de periodos académicos para futuras investigaciones, en específico, para mejorar la precisión y comprobar la certitud de los análisis realizados en este trabajo de titulación, se recomienda tratar de obtener una mayor cantidad de periodos académicos pero que incluyan los utilizados en esta investigación.
- Se recomienda para el investigar, que, durante la labor de recolección de datos, se pregunte en ese instante las descripciones de estos, es decir, se recomienda tener sumamente claro el contenido de la información que se obtiene, por lo que, se recomienda fuertemente que se dedique un buen porcentaje del tiempo a la fase de comprensión de los datos.

BIBLIOGRAFÍA

Agacino, R. (s. f.). *Conceptos fundamentales de estadística*. Retrieved from Cepal:

<https://www.cepal.org/sites/default/files/presentations/2.2-conceptos-fundamentales-estadisticas-ambientales.pdf>

Alteryx. (s. f.). *Exploración de datos*. Retrieved from Alteryx:

<https://www.alteryx.com/es/glossary/data-exploration>

Amazon. (n.d.). *¿En qué consiste la limpieza de los datos?* Retrieved from aws:

<https://aws.amazon.com/es/what-is/data-cleansing/>

Amazon. (n.d.). *¿Qué es una red neuronal?* Retrieved from AWS:

<https://aws.amazon.com/es/what-is/neural-network/#:~:text=Una%20red%20neuronal%20es%20un,lo%20hace%20el%20cerebro%20humano.>

Amazon. (s. f.). *¿Qué es la regresión logística?* Retrieved from aws:

<https://aws.amazon.com/es/what-is/logistic-regression/>

aprendeja. (n.d.). *La curva ROC (característica de funcionamiento del receptor)*. Retrieved from

aprendeja: <https://aprendeia.com/curvas-roc-y-area-bajo-la-curva-auc-machine-learning/>

Ayuware. (s. f.). *Cómo estandarizar datos y por qué es importante*. Retrieved from Ayuware:

<https://www.ayuware.es/blog/como-estandarizar-datos/>

Banco Mundial. (2020, febrero 7). *La importancia de proteger la “privacidad” en la era de los datos digitales*. Retrieved from Banco Mundial: <https://blogs.worldbank.org/es/voces/la-importancia-de-proteger-la-privacidad-en-la-era-de-los-datos-digitales>

Banco Pichincha. (2022, diciembre 10). *Banco Pichincha*. Retrieved from <https://www.pichincha.com/portal/blog/post/ley-proteccion-datos-ecuador-que-es#:~:text=El%20objetivo%20de%20esta%20ley,obligaciones%20y%20mecanismos%20de%20tutela%E2%80%9D>.

Borja Naranjo, G., Martínez Benítez, J., Barreno Freire, S., & Garo Jácome, O. (2021). *Factores asociados al rendimiento académico: Un estudio de caso* (Vol. vol. 23). REVISTA EDUCARE. Retrieved from <http://portal.amelica.org/ameli/jatsRepo/375/3752842004/index.html>

Carron, J. (2021, diciembre 12). *Violin Plots 101: Visualizing Distribution and Probability Density*. Retrieved from Mode: <https://mode.com/blog/violin-plot-examples/#:~:text=What%20is%20a%20violin%20plot%3F,the%20density%20of%20each%20variable>.

Category Encoders. (2022). *Category Encoders*. Retrieved from Category Encoders: https://contrib.scikit-learn.org/category_encoders/

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (1999). *CRISP-DM 1.0*. SPSS.

Colombia DANE DIRPEN. (2014, Agosto). *Lineamientos para la Anonimización de microdatos*. Retrieved from <https://pdfslide.tips/documents/lineamientos-para-la-anonimizacion-de-microdatos.html>

Constitución de la República del Ecuador. (2008). artículo 66.

Dávila Cárdenes, N., García Artiles, M., Pérez Sánchez, J., & Gómez Déniz, E. (2015, septiembre 1). *Un modelo de regresión logística asimétrico que puede explicar la probabilidad de éxito en el rendimiento académico*.
doi:<https://doi.org/10.6018/rie.33.1.178481>

Educación 3.0. (n.d.). *Por qué el vínculo entre estudiantes y docentes es clave para aprender*. Retrieved from Educación 3.0: <https://www.educaciontrespuntocero.com/noticias/por-que-vinculo-estudiantes-docentes-es-clave-aprender/>

Espejo, L. M. (s. f.). *Diagrama de Caja y Bigotes*. Retrieved from utalca: http://inst-mat.utralca.cl/tem/sitiolmde/primeroguias-liceo/recuperacion/Diagrama_de_Caja_y_Bigotes-2.pdf

Gandhi, R. (2018, junio 7). *Support Vector Machine — Introduction to Machine Learning Algorithms*. Retrieved from towardsdatascience: <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>

González, J. J., & Pérez, J. C. (2008, octubre). *LOS ORÍGENES DEL FRACASO ESCOLAR EN ESPAÑA.UN ESTUDIO EMPÍRICO*. Retrieved from ResearchGate: https://www.researchgate.net/publication/28294142_Los_origenes_del_fracaso_escolar_en_Espana_Un_estudio_empirico

Google. (n.d.). *Conceptos básicos sobre la información personal identificable en los contratos y las políticas de Google*. Retrieved from Google Analytics: <https://support.google.com/analytics/answer/7686480?hl=es#:~:text=Google%20consider>

a%20informaci%C3%B3n%20personal%20identificable,Direcciones%20de%20correo%20postal

Google. (s. f.). *¿Qué es el aprendizaje automático?* Retrieved from Google Cloud:

<https://cloud.google.com/learn/what-is-machine-learning?hl=es-419>

Guía para tratamiento de datos personales en administración pública. (2019, agosto 15).

Retrieved from Acuerdo Ministerial 12: <https://www.gobiernoelectronico.gob.ec/wp-content/uploads/2019/11/Gu%C3%ADa-de-protecci%C3%B3n-de-datos-personales.pdf>

Guizado Salazar, G., Valenzuela Rodríguez, M., & Vallejo Quispe, P. (2020). *Desempeño docente y el rendimiento académico de los estudiantes de la Facultad de Tecnología en la Universidad Nacional de Educación de Perú*. Conrado.

doi:http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1990-86442020000100200

Hernández, J. (2022, septiembre 22). *Inteligencia artificial: qué aporta y qué cambia en el mundo del trabajo*. Retrieved from Inter-American Development Bank:

<https://blogs.iadb.org/trabajo/es/inteligencia-artificial-que-aporta-y-que-cambia-en-el-mundo-del-trabajo/>

Hotz, N. (2023, enero 19). *What is CRISP DM?* Retrieved from Data Science Process Alliance:

<https://www.datascience-pm.com/crisp-dm-2/>

Huerta Reyes, M., Pérez García, L., López González, J., Uscanga Hermida, R., & Ríos Carrillo, P. (2023, enero). *Artículo Estrés y rendimiento académico 2*.

doi:10.25009/revmedforense.v8i1.2981

Ibarrera. (2020, octubre 12). *El temor a los datos duplicados – Una guía para la deduplicación de datos*. Retrieved from Data Ladder: <https://dataladder.com/es/el-temor-a-los-datos-duplicados-una-guia-para-la-deduplicacion-de-datos/#:~:text=Los%20datos%20duplicados%20son%20un,empresa%20a%20tomar%20decisiones%20err%C3%B3neas>.

IBM. (n.d.). *¿Qué es un árbol de decisión?* Retrieved from IBM: <https://www.ibm.com/es-es/topics/decision-trees>

IBM. (2021, agosto 17). *Conceptos básicos de ayuda de CRISP-DM*. Retrieved from ibm: <https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview>

IBM. (2022, septiembre 13). *Regresión Logística*. Retrieved from IBM: <https://www.ibm.com/docs/es/spss-statistics/saas?topic=regression-logistic>

IBM. (s. f.). *Conceptos básicos sobre comprensión de datos*. Retrieved from <https://www.ibm.com/docs/es/spss-modeler/saas?topic=understanding-data-overview>

IBM. (s.f.). *¿Qué es Machine Learning?* Retrieved from IBM: <https://www.ibm.com/mx-es/analytics/machine-learning>

IBM. (n.d.). *What is random forest?* Retrieved from IBM: <https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems>.

jmp. (s. f.). *Análisis exploratorio de datos*. Retrieved from jmp Estatistical Discovery: https://www.jmp.com/es_co/statistics-knowledge-portal/exploratory-data-analysis.html

jmp. (s. f.). *Introducción a la estadística*. Retrieved from JMP: Histograma. (s. f.). Introducción a la estadística | JMP. https://www.jmp.com/es_co/statistics-knowledge-portal/exploratory-data-analysis/histogram.html

Lane, D. M. (s. f.). Graphing Qualitative Variables. In D. Scott, M. Hebl, R. Guerra, D. Osherson, & H. Zhimmer, *Online Statistics Education: An interactive Multimedia Course of Study* (p. 692). David M. Lane. Retrieved from https://onlinestatbook.com/Online_Statistics_Education.pdf

Loayza, V., & Loayza, X. (s. f.). *Por Qué Es Importante Identificar El Tipo de Los Datos*. Retrieved from Scribd: <https://es.scribd.com/document/483227513/Por-que-es-importante-identificar-el-tipo-de-los-datos-acv01#>

López, F. J. (2022, noviembre 15). *Apuntes de estadística*. Retrieved from bioestadistica.uma.es: <https://www.bioestadistica.uma.es/apuntesMaster/exploraci%C3%B3n-de-los-datos.html>

Lucidchart. (n.d.). *Qué es un diagrama de árbol de decisión*. Retrieved from Lucidchart: <https://www.lucidchart.com/pages/es/que-es-un-diagrama-de-arbol-de-decision>

Martínez Benítez, J., Borja Naranjo, G., Haro Jácome, O., & Barreno Freire, N. (2020). *Factores personales, sociales e institucionales asociados al rendimiento académico en la educación superior* (Vol. Vol. 4.(3)). MQRInvestigar.

Ordóñez, L. (2007). *El desarrollo tecnológico en la historia*. Lima: Areté.

Orellana, L. (2001, marzo 1). *Estadística Descriptiva*. Retrieved from https://www.dm.uba.ar/materias/estadistica_Q/2011/1/modulo%20descriptiva.pdf

pandas. (s. f.). *pandas.DataFrame.describe*. Retrieved from pandas:

<https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.describe.html>

pandas. (s. f.). *Python Data Analysis Library*. Retrieved from <https://pandas.pydata.org/>

PowerData. (2023, mayo 23). *Enmascaramiento de datos. ¿Qué es? ¿Qué debo tener en cuenta?*

Retrieved from <https://www.powerdata.es/enmascaramientode-datos>

Reglamento General Europeo de Protección de Datos Personales. (2018).

Rico Páez, A., & Gaytán Ramírez, N. (2022, abril 20). *Modelos predictivos del rendimiento académico a partir de características de estudiantes de ingeniería*.

doi:https://doi.org/10.33010/ie_rie_rediech.v13i0.1426

Rueda, J. F. (2019, noviembre 4). *CRISP-DM: una metodología para minería de datos en salud*.

Retrieved from healthdataminer.com: <https://healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud/>

Ruíz Dodobara, F. (2005). *Influencia de la autoeficacia en el ámbito académico*. Revista Digital de Investigación en Docencia Universitaria. doi:<https://doi.org/10.19083/ridu.1.33>

scikit-learn. (n.d.). *sklearn.ensemble.RandomForestClassifier*. Retrieved from scikit-learn:

[https://scikit-](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html)

[learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html)

scikitlearn. (n.d.). *sklearn.preprocessing.MinMaxScaler*. Retrieved from scikitlearn:

[https://scikit-](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html)

[learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html)

scikitlearn. (n.d.). *sklearn.preprocessing.StandardScaler*. Retrieved from scikitlearn:

[https://scikit-](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html)

[learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html)

TIBCO. (n.d.). *¿Qué es el aprendizaje supervisado?* Retrieved from TIBCO:

<https://www.tibco.com/es/reference-center/what-is-supervised-learning>

Umbarilla, M. A. (2022, abril 6). *Gráfico de Densidad*. Retrieved from RPubS by RStudio:

<https://rpubs.com/Alejandra01/887394>

UNAM. (s. f.). *Identificación de variables en una investigación*. Retrieved from

[http://uapas1.bunam.unam.mx/matematicas/iden_variables/#:~:text=Esto%20permite%20saber%20al%20investigador,dependiente\)%20con%20los%20cambios%20provocados.](http://uapas1.bunam.unam.mx/matematicas/iden_variables/#:~:text=Esto%20permite%20saber%20al%20investigador,dependiente)%20con%20los%20cambios%20provocados.)

Unión Europea. (2016). *Reglamento General de Protección de Datos*.

United Kingdom. Information Commissioner's Office. (2017, febrero 1). *Anonymisation:*

managing data protection risk code of practice. Retrieved from

<https://ico.org.uk/media/1061/anonymisation-code.pdf>

Universidad de Hawaii. (n.d.). *Practices of Science: Precision vs. Accuracy*. Retrieved from

Exploring our fluid earth:

[https://manoa.hawaii.edu/exploringourfluidearth/physical/world-ocean/map-](https://manoa.hawaii.edu/exploringourfluidearth/physical/world-ocean/map-distortion/practices-science-precision-vs-accuracy#:~:text=Precision%20and%20accuracy%20are%20two,Precision%20is%20independent%20of%20accuracy.)

[distortion/practices-science-precision-vs-](https://manoa.hawaii.edu/exploringourfluidearth/physical/world-ocean/map-distortion/practices-science-precision-vs-accuracy#:~:text=Precision%20and%20accuracy%20are%20two,Precision%20is%20independent%20of%20accuracy.)

[accuracy#:~:text=Precision%20and%20accuracy%20are%20two,Precision%20is%20independent%20of%20accuracy.](https://manoa.hawaii.edu/exploringourfluidearth/physical/world-ocean/map-distortion/practices-science-precision-vs-accuracy#:~:text=Precision%20and%20accuracy%20are%20two,Precision%20is%20independent%20of%20accuracy.)

- Vargas, G. M. (2013, julio 31). *Factores asociados al rendimiento académico en estudiantes universitarios desde el nivel socioeconómico: Un estudio en la Universidad de Costa Rica*. Retrieved from Revista Electrónica Educare:
[https://www.scielo.sa.cr/scielo.php?script=sci_arttext&pid=S1409-42582013000300004#:~:text=El%20rendimiento%20acad%C3%A9mico%20en%20estudiantes,et%20al.%2C%202002\).](https://www.scielo.sa.cr/scielo.php?script=sci_arttext&pid=S1409-42582013000300004#:~:text=El%20rendimiento%20acad%C3%A9mico%20en%20estudiantes,et%20al.%2C%202002).)
- Véliz, C. (2020). *Aprendizaje automático. Introducción al aprendizaje profundo*. Perú: El Fondo Editorial de la Pontificia Universidad Católica del Perú.
- Ziemer, H. (s. f.). Variables. In D. Scottt, M. Hebl, R. Guerra, D. Osherson, & H. Zimmer, *Online Statistics Education: An Interactive Multimedia Course of Study* (p. 692). David M. Lane. Retrieved from onlinestatbook:
https://onlinestatbook.com/Online_Statistics_Education.pdf
- Zumárraga Espinosa, M., & Cevallos Pozo, G. (2022, julio 26). *Autoeficacia, procrastinación y rendimiento académico en estudiantes universitarios de Ecuador*.
doi:<https://doi.org/10.17163/alt.v17n2.2022.08>

ANEXOS

Anexo 1: Matriz de evaluación de Información Personal Identificable.xlsx

Anexo 2: Resultado de los códigos para la identificación de variables únicas en los reportes.pdf

Anexo 3: datos_para_el_investigador.csv

Anexo 4: index.html

Anexo 5: prediccion.html

Anexo 6: resultado.html

Anexo 7: modelo_tesis.h5