PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

Predicción de la deserción estudiantil, mediante métricas y técnicas de minerías de datos, en el Instituto Superior Tecnológico Los Andes (ISTLA).

Trabajo de Titulación previo a la obtención del título de Magister en Sistemas de Información mención Data Science

Línea de Investigación: Tecnologías de la información y la comunicación.

Autor:

JAVIER JOSÉ CEVALLOS FARÍAS

Director:

Dr. Rafael Melgarejo

Quito – Ecuador

Octubre, 2022

PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

HOJA DE APROBACIÓN

PREDICCIÓN DE LA DESERCIÓN ESTUDIANTIL, MEDIANTE MÉTRICAS Y TÉCNICAS DE MINERÍAS DE DATOS, EN EL INSTITUTO SUPERIOR TECNOLÓGICO LOS ANDES (ISTLA).

Línea de Investigación: Tecnologías de la información y la comunicación**.**

Autor:

JAVIER JOSÉ CEVALLOS FARÍAS

Rafael Melgarejo, Dr.
DIRECTOR DE TRABAJO DE TITULACIÓN

Santo Domingo – Ecuador

Octubre, 2022

DECLARACIÓN DE AUTENTICIDAD

Yo, JAVIER JOSÉ CEVALLOS FARÍAS portador de la cédula de ciudadanía No. 171377802-3, afirmo que el presente trabajo de investigación que presento como proyecto de, previo la obtención del Título de Magister en Sistemas de Información mención Data Science, son autoría originales y personales.

Por lo expuesto declaro que todo el contenido como argumentos conclusiones y referencias que generen algún problema legal son responsabilidad mía .

JAVIER JOSÉ CEVALLOS FARÍAS

CI. 171377802-3

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE ANEXOS

RESUMEN

La deserción estudiantil en niveles de educación superior en la actualidad se ha convertido en un problema a nivel económico, social, comunitario y educativo a nivel mundial, puesto que una persona que no se encuentra con los conocimientos necesarios para enfrentar al entorno que le rodea tampoco le será posible ser competitivo y por lo tanto podría quedar rezagado del sistema económico, social, laboral, lo que estaría en detrimento de su calidad de vida. El presente proyecto tiene como objetivo determinar los factores de la deserción de los estudiantes del Instituto Superior Tecnológico Los Andes comprendida entre los periodos 2019 hasta el 2022. La metodología a utilizar es de tipo mixta, es decir cualitativa y cuantitativa, ya que por una parte se realizó un análisis objetivo del tema en cuestión y también resultados de la toma de datos con tablas de frecuencia, gráficos estadísticos, que detallan de forma lógica los factores de la deserción estudiantil en los sujetos de estudio. Se espera que este estudio sirva para identificar los factores que determinan la deserción estudiantil para posterior crear estrategias y tomar decisiones utilizando la minería de datos ya que se permitirá descubrir patrones de comportamiento a partir de un gran conjunto de datos. Y de esta manera aportar concientizando a la población estudiantil actual sobre la importancia del estudio para su vida profesional y personal pues una sociedad con educación es una sociedad con esperanza de obtener mejor estabilidad económica, laboral, social, y familiar.

Palabras clave: Deserción, estudiantes, metodología, factores, socioeconómico.

ABSTRACT

Student desertion in higher education levels today has become a problem at an economic, social, community and educational level worldwide, since a person who does not have the necessary knowledge to face the environment that surrounds him or her it will be possible for him to be competitive and therefore he could be left behind in the economic, social, labor system, which would be detrimental to his quality of life. The objective of this project is to determine the factors of the desertion of the students of the Los Andes Higher Technological Institute between the periods 2019 to 2022. The methodology to be used is of a mixed type, that is, qualitative and quantitative, since on the one hand An objective analysis of the subject in question will be carried out and data collection results will also be included with frequency tables, statistical graphs, which logically detail the factors of student dropout in the study subjects. It is expected that this study will serve to identify the factors that determine student dropout in order to later create strategies and make decisions using data mining, since it will allow the discovery of behavior patterns from a large set of data. And in this way, contribute by making the current student population aware of the importance of studying for their professional and personal life, since an educated society is a society with the hope of obtaining better economic, labor, social, and family stability.

Keywords: Dropout, students, methodology, factors, socioeconomic.

1. INTRODUCCIÓN

1.1. Planteamiento del problema

La deserción estudiantil es uno de los fenómenos sociales y educativos a nivel mundial que afecta a todas las instituciones de educación, particularmente en la educación superior, en Europa siendo países desarrollados han venido teniendo un bajo índice de abandono de estudios superiores pero quedando evidenciado que existe el problema, siendo España uno de los países que tiene la mayor tasa de deserción escolar en la Unión Europea, según Eurostat afirma que el 17,3% de los españoles que tienen un rango de edad entre 18 y 24 años no han considerado continuar formándose después de finalizar su secundaria en 2019 (Álvarez, 2020).

Un estudio realizado por Espíndola & León (2002) Revista Iberoamérica de Educación, demuestra que a:

A nivel de Latinoamérica la deserción estudiantil se marca en el aspecto económico y social debido a la escasa capacidad de retención de los niños y adolescentes en las escuelas. Esto significa que la gran mayoría de los niños y niñas no completan ese ciclo muchos de ellos ni siquiera terminan la primaria y los pocos que llegan a niveles superiores optan por abandonar sus estudios por la carencia de ingresos económicos a los hogares y otros factores relacionados con estratos pobres, esto dificulta la política pública de justicia social e igualdad de oportunidades para todos y todas si se compara con los estratos de un nivel socio-económico medio y alto, por lo cual, es evidente que las condiciones socioeconómicas son decisivas cuando se habla de formación humana para la reproducción de la desigualdad social.

Asimismo, resulta inquietante que, en la mayor parte de zonas urbanas exista un elevado porcentaje que sobrepasa el 50% de los estudiantes que dejan sus estudios de escuela sin culminarlos por la carencia de recursos en sus hogares. Más aún, en Argentina (Gran Buenos Aires y total urbano), Chile, Costa Rica, Honduras, México, Panamá, Paraguay (Asunción y Departamento Central) y Uruguay, el 60% o más de los niños que se retiran en el transcurso de la primaria se concentran en el 25% de los hogares más pobres.

Ecuador no está ajeno ante esta situación que se ve afectado por muchos obstáculos como la corrupción y la pobreza dando un alto índice de deserción universitaria.

La provincia Tsáchila cuenta con universidades e institutos que sufren esta problemática. Es así que el Instituto Superior Tecnológico Los Andes siendo una institución particular de educación superior tiene estos problemas de deserción estudiantil en sus diferentes carreras, para lo cual se va a analizar las variables causantes a la problemática como, economía, académica, inadecuado selección de carrera, entre otras.

En el Instituto Superior Tecnológico "Los Andes" por lo menos el 15% de los estudiantes ha optado por abandonar los estudios, según la entrevista no formal realizada al rector de la institución indicó que esto se debe a que en su mayoría al ser una institución educativa superior particular los estudiantes no pueden cumplir con los pagos que la misma exige, sumando a ello que gran parte de los estudiantes reciben ayuda económica de sus padres para continuar con sus estudios, mientras que otra parte son jefes de hogar y en consecuencia se ven en la opción de abandonar sus estudios, por lo que se incluye el documento a dicho segmento.

La presente investigación se la realizó en el Instituto Superior Tecnológico "Los Andes" (ISTLA), involucrando a toda la comunidad educativa.

1.2. Preguntas de la investigación

Pregunta general.

¿Cómo predecir la deserción de estudiantes de la carrera de Sistemas del Instituto Superior Tecnológico Los Andes (ISTLA)?

De donde se generan las preguntas específicas.

- ¿Cuáles son las características socioeconómicas de los estudiantes del Instituto Superior Tecnológico Los Andes (ISTLA)?
- ¿De qué manera aplicar métricas estadísticas para analizar las variables que ocasionan la deserción estudiantil?
- ¿Cómo aplicar técnicas de minería de datos para predecir la deserción estudiantil?
- ¿De qué manera interpretar los resultados de la predicción estudiantil?

1.3. Objeto de estudio

Comportamiento de los estudiantes.

1.4. Campo de acción

Estudiantes del ISTLA

1.5. Justificación

El presente proyecto incluye el análisis de un problema no solo a nivel educativo sino en el campo social e incluso familiar como lo es la deserción estudiantil, entendida como tal al abandono de cualquier nivel de educación en un periodo específico antes de haberlo completado o culminarlo en su totalidad.

Es así que el propósito fundamental de este proyecto es identificar los variables influyentes en la deserción estudiantil, dado que es un problema que se percibe cada vez con mayor frecuencia y que sobre todo ha terminado por debilitar a la sociedad, pues una sociedad sin educación también trae como consecuencia la exclusión social, la poca información y especialización, así como la posibilidad de no ir a la par con las exigencias competitivas laborales de la actualidad, lo que repercute en el desarrollo e inserción laboral futura y por lo tanto va en detrimento de la calidad de vida de las personas.

Con los resultados de este proyecto se busca concientizar tanto a autoridades de educación como a docentes, padres y estudiantes sobre la urgencia por crear planes y programas que permitan motivar a la culminación de la vida académica teniendo claro que esto llevará a estas personas a aspirar una mejor calidad de vida para ellos y su familia.

Los beneficiarios directos son los estudiantes, quienes de alguna manera o situación familiar, afectivo, emocional, entre otros podrían estar pensando desertar de sus estudios, lo cual es un llamado a la concientización para evitar y frenar esta ola de deserción estudiantil en la actualidad, otros beneficiarios son los padres quienes podrán tener la satisfacción y tranquilidad de que sus hijos (as) tengan conciencia de la importancia del estudio y de la sociedad en general pues a mayor nivel de estudios mayor será la calidad de vida de las personas.

1.6. Objetivos de la investigación

1.6.1. Objetivo general.

Predecir el índice de deserción de estudiantes mediante métricas y técnicas de minería de datos de la carrera de Sistemas del Instituto Superior Tecnológico Los Andes (ISTLA).

1.6.2. Objetivos Específicos

- Determinar las características socioeconómicas de los estudiantes del Instituto Superior Tecnológico Los Andes (ISTLA)
- Aplicar métricas estadísticas para analizar las variables que ocasionan la deserción estudiantil.
- Aplicar técnicas de minería de datos para predecir la deserción estudiantil
- Evaluar los resultados de la predicción de la deserción de estudiantes mediante técnicas de minería de datos.

El trabajo de titulación está estructurado de la siguiente forma: Introducción, Revisión de la literatura donde se argumenta científicamente las variables de la investigación, metodología que se emplea para obtener los resultados, conclusiones, recomendaciones, referencias bibliográficas y anexos que respalda la investigación.

2. REVISIÓN DE LA LITERATURA

2.1. Fundamentos teóricos

Existe el documento denominado “El Problema de la Deserción Escolar en la Producción Científica Educativa” (Hernández y Aranda, 2017). Este artículo incluye un análisis exploratorio de la producción científica en español sobre la deserción escolar, que se publicó entre el año 2000 al 2016 desde una aproximación bibliométrica. El fundamento de datos usados fue de Dialnet, donde se obtuvieron 53 documentos. Se verifica que los resultados indicaron que existe una repercusión de la deserción educativa, y se propone el establecer relaciones universitarias, de tal forma que sea sólida la publicación para la mayor divulgación y conciencia de la importancia del estudio.

También se encontró el documento que se denomina “Causas y Consecuencias de la Deserción escolar en el bachillerato: caso Universidad Autónoma de Sinaloa” Ruiz, García, y Pérez (2016). El fin de la investigación es el definir las causas y consecuencias en cuanto a aspectos personal, económico, social, que produce deserción escolar de estudiantes de preparatoria, en la UAS, se hizo un estudio en escuelas que pertenecen al municipio de Fuerte, Sinaloa, en la unidad académica San Glas y las extensiones de esta que son la Constancia y las Higueras. Se dio paso a una metodología mixta la muestra se tomó con 18 desertoras, 17 desertores de ciclo escolar, 3 profesoras y 7 profesores, 12 alumnas que no desertaron, 2 directivas y 2 directivos. Los resultados arrojaron que el factor clave de deserción escolar es el personal, y se destaca aspectos como el no aprobar materias o el casarse, siendo que la principal consecuencia es de tipo económica y se destaca el círculo de pobreza.

También está el documento denominado “La problemática actual de la deserción escolar, un análisis desde lo local” Martínez y Ortega (2020), se habla de la educación como un activo para toda persona y la sociedad, por lo que se dice que la mayor deserción escolar trae problemas de extra edad, rezagos educativos para poblaciones en edad de cursar la educación básica en el siguiente ciclo. Se considera que una causa es el inicio de consumo de drogas y alcohol por los estudiantes. Como conclusión se llegó a la necesidad de concientizar y que el estado cree programas para concientizar a la población en edad estudiantil sobre la no deserción académica.

2.2. Marco Conceptual

2.2.1. Deserción estudiantil

En el entorno internacional se han realizado varios proyectos de investigación en los que se ha aplicado la minería de datos para el descubrimiento de factores de deserción estudiantil, en el ámbito colombiano en la Universidad de Nariño y la Institución Universitaria CESMAG de la ciudad de Pasto (Colombia) según Jiménez y Timaràn (2015) se utilizó técnica de minería de datos:

Al aplicar etapas de pre procesamiento y transformación con la finalidad de obtener conjuntos de datos limpios y aplicando técnicas de clasificación basadas en arboles de decisión, asociación y clustering, utilizando la herramienta libre de minería de datos Weka, los resultados determinaron que un factor para la deserción estudiantil ha sido la obtención de un promedio bajo en notas y materias perdidas en los primeros semestres. (p. 447)

Del mismo modo, información recabada de instituciones como la Universidad Católica del Norte de Chile en ingeniería han determinado que:

> Las variables que mejor explican la deserción de un estudiante son, las razones socioeconómicas y el puntaje de ingreso a la universidad. Según el árbol de decisión construido se concluye que la retención se sitúa en un 78,3%. La calidad de los clasificadores permite asegurar que sus predicciones son correctas, con niveles estadísticos de curva ROC de 76%, 75% y 83% de acierto para los clasificadores de red bayesiana, árbol de decisión y red neuronal respectivamente. (Miranda y Guzmán, 2017, p. 61)

De igual forma, desde la misma línea de investigación se encuentra un estudio elaborado por la Universidad Arturo Prat en el mismo contexto, donde se plantea que:

> Se utilizó CRISP-DM como metodología para guiar las etapas del proyecto y se analizaron tres diferentes modelos de clasificación: árboles de decisión, métodos bayesianos y redes neuronales, con el fin de evaluar su comportamiento, encontrándose que Random Forest es el algoritmo de mejor desempeño general, con un 88,9% de exactitud, mientras que el algoritmo Naive Bayes resulto ser el más adecuado

> para dar respuesta a los objetivos del negocio, dados los niveles de sensibilidad alcanzados. (Torres, et al., 2016, p. 73)

2.2.2. Definición de deserción

La UNESCO (2020) define la deserción estudiantil como "como el porcentaje de estudiantes que, habiendo estado matriculado en un año, deja de estudiar y no vuelve a matricularse en el siguiente año". Según Muñoz (2013) define la deserción como "el abandono de un estudiante del sistema escolar por distintas variables agrupadas como socioeconómicas, institucionales y académicas".

Ante esta problemática se ha buscado soluciones por medio de herramientas que permitan una predicción sobre algunos rasgos del alumno que pueda desertar y acercarse a las posibilidades que abandone completamente sus estudios. Para esta investigación las técnicas usadas con mayor alcance son de Machine Learning y Minería de Datos.

2.2.3. Técnicas de aprendizaje automático y minería de datos

Se podría definir al aprendizaje automático como la programación de equipos de cómputo utilizando ciertos parámetros para que se optimice el rendimiento. Existen diversos modelos para ello, puede ser de forma predictivas para determinar inferencias o simplemente descriptivo para tener un panorama de los datos. (Ethem, 2014). No existe una mínima cantidad de algoritmos dedicados al aprendizaje automático, sino que hay muchos creados con ese fin. Autores como Ethem (2014) presentan taxonomías de técnicas a partir del proceso de aprendizaje que se dividen en: con supervisión, sin supervisión, cuasi-supervisada y de refuerzo para el aprendizaje.

La aplicación de técnicas de minería de datos se puede emplear en diferentes situaciones en función del método que se elija, mismos que tienen diversas clasificaciones, entre ellos se encuentran de clasificación, analíticos de asociación y agrupamiento (Vieira, et al., 2009). Es preciso relatar que dicha clasificación está directamente relacionada con el aprendizaje con supervisión, sin embargo, también existen técnicas vinculadas al aprendizaje sin supervisión que son: analíticas de asociación y de agrupamiento, la primera ayuda a presentar vínculos escondidos en grandes conjuntos de datos y la segunda divide la agrupación de datos en grupos que resulten significativos, de esta forma si se encuentra una significatividad en los grupos, las conjuntos deben captar la estructura propia de los datos.

Finalmente, entre los métodos de clasificación y aprendizaje guiado se puede apreciar a las redes neuronales artificiales, a los árboles de decisión, la regresión logística, equipos de soporte vectorial y procedimientos de ensamblaje de dichos algoritmos.

2.2.4. Arboles de decisión (CHAID)

Con la finalidad de reducir la tasa de deserción estudiantil, y mejorar la tasa de retención de los estudiantes, se ve la necesidad de desarrollar un sistema o método que ayude a determinar las situaciones de riesgo académico de los estudiantes, y como consecuencia emplear medidas oportunas para su retención. En estudio realizado a los estudiantes de la Universidad Nacional de San Agustín (UNSA), se utilizó la técnica de árboles de decisión CHAID, utilizada para segmentar, estratificar, predecir, reducir datos y filtrar variables, etc. Se define CHAID como:

Un algoritmo para la construcción de árboles de decisión basado en el testeo de significancia ajustada, explora datos de forma rápida y eficaz, y crea segmentos y perfiles relacionados con el resultado deseado. Usa la chi-cuadrado para medir el grado de correlación entre las variables independientes y la clase. (Bedregal, Aruquipa y Cornejo, 2020, p. 595)

2.2.5. Modelado de datos

Domínguez (2018) afirma que el modelado de datos es “el proceso de documentar un diseño de sistema de software complejo como un diagrama de fácil comprensión, usando texto y símbolos para representar la forma en que los datos necesitan fluir” (p. 33).

2.2.6. Analítica Académica

Referente al concepto de analítica académica Norris & Lefrere, (como se citó en Contreras, Rodríguez, y Fuentes, 2021) definen a la analítica académica como los diversos procesos de evaluación y análisis de datos permitiendo medir, mejorar y comparar el desempeño de individuos, programas o instituciones.

2.2.7. Ética y educación

La capacidad para ejercer como profesionales debe acarrear habilidades adquiridas a lo largo de sus años, en especial en la juventud dado que el proceso de aprendizaje estará rodeado de situaciones que ayuden y promuevan el desarrollo de habilidades para una óptima vida académica y social.

Para Santos, Mella y García (2021) “los profesores deben de ser conscientes que su trabajo tiene que ver con la construcción de buenos profesionales, sin olvidar que, primordialmente deben ser ciudadanos responsables a la altura del título obtenido” (p 167).

Entorno al ámbito ético en la educación y la formación moral que se puede brindar a los alumnos dentro de las aulas educativas:

> El acto moral es una praxis, una actividad humana en la que se funden el fin y los medios. La tekné, según Aristóteles, es otra forma de actividad cuya finalidad es externa a los medios: (…) La praxis, como actividad ética, contribuye al desarrollo del agente moral. La praxis es una auto práctica para la cual el agente moral establece como condición para volverse más lúcido y libre a través de la acción. (Moreau, 2019, p.32)

3. METODOLOGÍA DE LA INVESTIGACIÓN

3.1. Metodología

La presente investigación incluirá una metodología mixta es decir cualitativa y cuantitativa, ya que por una parte se analiza el objeto de estudio en base al criterio objetivo del investigador que en este caso es la deserción estudiantil y por otro lado se realizará un estudio cuantitativo debido a que se expondrán los hallazgos encontrados de la toma de instrumentos como la encuesta y se emitirá las correspondientes tablas de frecuencia, gráficos, figuras que son parte de la investigación.

3.1.1. Método

El método a utilizar *es CRISP-DM (Cross-Industry Standard Process for Data Mining)* como metodología de extracción de conocimiento ya que se ajusta a diferentes realidades del proyecto generando la documentación necesaria del proceso con la finalidad que éste sea replicable.

Esta metodología plantea el ciclo de vida de un proyecto de minería de datos que consta de seis etapas que comienza con una buena comprensión y conocimiento del negocio y la necesidad del proyecto, y concluye con el despliegue de la solución que cumple la necesidad específica del negocio. (Joyanes, 2015, p.300)

1. Comprensión del negocio.
2. Comprensión de los datos.
3. Preparación de los datos.
4. Modelado (construcción de modelado)
5. Evaluación.
6. Despliegue (desarrollo).

3.1.1.1. Comprensión del negocio

Se trata de una fase que abre el proceso y se encuentra enfocada a la comprensión de los propósitos y exigencias del proyecto que se plantea, desde la construcción del negocio, "implica acceder a los datos y explorarlos con la ayuda de tablas y gráficos que se pueden organizar en IBM® SPSS Modeler utilizando la herramienta de proyectos CRISP-DM" (IBM, 2021, p.1).

3.1.1.2. Comprensión de los datos

La fase de comprensión de datos de CRISP-DM "implica estudiar más de cerca los datos disponibles de minería. Siendo este el paso esencial para evitar problemas inesperados durante la siguiente fase (preparación de datos) que suele ser la fase más larga de un proyecto" (IBM, 2021, p 2).

La eficacia de la data posee algunas dimensiones como: la exactitud (que muestre lo sucede), totalidad (que los datos se hallen completos dentro del sistema), oportunidad (que se pueda acceder a ellos siempre que se necesite), relevancia, nivel de detalle y consistencia (que no haya variación de la data en todos los ámbitos y sistemas), por consiguiente, es indispensable que se represente la forma de los datos en cada uno de los ámbitos. (CEUPE,2022)

3.1.1.3. Preparación de los datos.

El objetivo de esta fase es "localizar los datos requeridos para el análisis, extraerlos de sistema de información y asegurar su calidad para el posterior análisis con las técnicas de minería de datos" (Aguirre, 2026, p.83).

> Las tareas de preparación o de limpieza de datos van a ser realizadas repetidas veces y no en cualquier orden. Entre estas tareas tenemos la selección de tablas, registros y atributos, así como la transformación y la limpieza de datos para su preparación para las herramientas de modelado. (CEUPE, 2022, p.1)

3.1.1.4. Modelado

Se seleccionan diversas técnicas de modelado adecuadas a un conjunto de datos ya preparado a fin de centrarse en las necesidades específicas del negocio. Para (Joyanes, 2019) Las subfases son:

En esta etapa se busca conseguir las siguientes cuatro operaciones:

- Selección de la técnica de modelado apropiada.
- Diseño de evaluación (generación de un diseño de comprobación: plan de pruebas)
- Construcción del modelo.
- Evaluación del modelo.

3.1.1.5. Evaluación

Evaluar el modelo de la fase anterior, comprobar si el modelo sirve para responder a las necesidades del negocio desde un punto de vista de analítica de data. Es decir, antes de que se haga la presentación final y la puesta en marcha, es necesario que se hagan las pruebas y la revisión de las etapas ejecutadas durante la elaboración del modelo, utilizando los criterios de rendimiento comercial establecidos en el inicio del proyecto, lo que servirá para comparar el modelo obtenido con los objetivos del negocio (IBM, 2021).

3.1.1.6. Despliegue

Según Joyanes (2019) la fase de despliegue "trata de explotar la potencialidad de los modelos, integrarlos en los procesos de toma de decisiones de la organización, difundir informes sobre el conocimiento extraído" (p 240).

Cuando la información encontrada sea presentada al beneficiario, las formas de evaluación pueden optimizarse, de esta forma, el proceso de minería puede refinarse y los datos generados permiten ser seleccionados o sirven para su transformación, también se pueden agregar nuevas fuentes de datos con la finalidad de obtener resultados diferentes o más convenientes.

3.1.2. Las herramientas a utilizar

- Excel
- Jupyter Notebook con lenguajes Phyton, Panda, dentro de estas las librerías necesarias para el desarrollo del proyecto.
- Google colab

3.2. Propuesta

Para este proyecto se utilizarán los datos del total de los estudiantes matriculados en el Instituto Superior Tecnológico “Los Andes” desde los periodos 2019 hasta el 2021. Para la generación de los modelos se procederá a una minería de datos en base a métricas y análisis estadísticos para la limpieza de la data.

Además, se pretende aplicar CRISP-DM para la generación del conocimiento y se pretende determinar los factores con mayor repetitividad presentes en los estudiantes, causantes de la deserción estudiantil, tomando en cuenta la situación socioeconómica y académica, en consecuencia, mejorar e implementar políticas y estrategias que utiliza el Instituto Tecnológico “Los Andes” (ISTLA) para conseguir una mejor retención de estudiantes y su permanencia en las carreras el ISTLA ha conformado estrategias empíricas, desde recursos humanos, con métodos matemáticos que son predictores de los estudiantes con riesgo de deserción, esto se almacena en un sistema denominado SIGALA que cruza información académica, financiera y psicológica y presenta una óptica clara de la situación de los estudiantes.

3.3. Aplicación de la metodología CRISP-DM

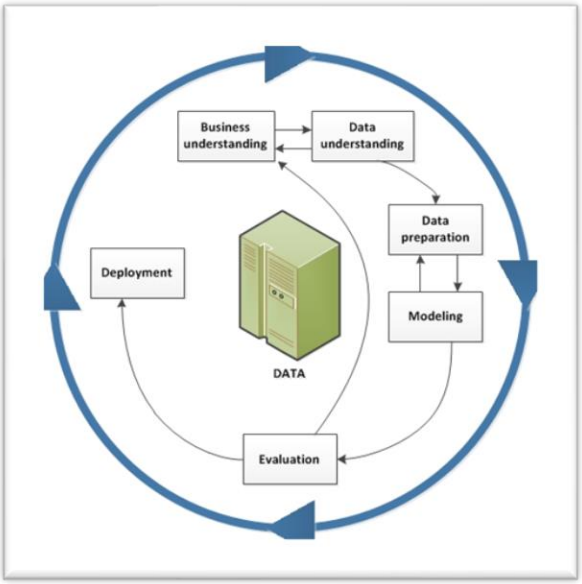


Figura: 1 Metodología CRISP-DM
Fuente: (IBM, 2021)

El modelo de CRISP-DM es flexible y se pueden personalizar fácilmente. Por ejemplo, si su organización intenta detectar actividades de blanqueo de dinero, es probable que

> necesite realizar una criba de grandes cantidades de datos sin un objetivo de modelado específico. En lugar de realizar el modelado, su trabajo se centrará en explorar y visualizar datos para descubrir patrones sospechosos en datos financieros. CRISP-DM permite crear un modelo de minería de datos que se adapte a sus necesidades concretas. (IBM, 2021, p.1)

3.3.1. Comprensión del negocio

En la actualidad, una de las problemáticas con mayor fuerza que se presentan en las Instituciones educativas a nivel Superior son los altos niveles de deserción, según la Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación (Senescyt) Citado por El Expreso (2019) señala que "existe un 26 % de estudiantes que han abandonado su carrera a inicios de la misma". Siendo uno de los principales factores que motivan a la deserción la situación socioeconómica acompañado del bajo rendimiento en las asignaturas.

Para reducir esta problemática en las instituciones educativas de nivel superior, se propone la creación y aplicación de modelos predictivos que basados en la implementación y utilización de herramientas de minería de datos permitan realizar un seguimiento académico de los estudiantes con el propósito de lograr identificar y predecir mediante un sistema de alertas anticipadas factores de riesgo que conlleven a la deserción del estudiante.

En instituciones educativas, como el Instituto Superior Tecnológico "Los Andes", se aplicaría el diseño de modelos predictivos con el propósito de crear y fomentar programas de prevención estratégicos para estudiantes con situaciones de riesgo socioeconómicos, rendimiento académico y evitar las futuras deserciones.

3.3.1.1. Objetivos del negocio

El Instituto Superior Tecnológico Los Andes, siendo una institución de educación superior tiene por objetivo principal el graduar a la mayor parte de sus educandos matriculados desde su primer nivel hasta el último, es así que los objetivos del proyecto están alineados a esa meta en conjunto, están encaminados a mejorar la toma de decisiones en cuestiones académicas y administrativas del instituto. En cuanto a los objetivos específicos se obtiene información relevante en cuanto a la situación socioeconómica y académica de los estudiantes, con utilización de métricas que ayudarán al entendimiento de la problemática.

3.3.1.2. Criterios de éxito del negocio

- Para que el presente proyecto tenga éxito se plantea que la predicción estudiantil debe ser con el menor grado de error, esto para mejorar la toma de decisiones pedagógicas y administrativas.

- Bridar ayuda a tiempo a los estudiantes mejorando el porcentaje de alumnos matriculados y alumnos graduados.

3.3.1.3. Evaluación de la Situación

La investigación está enmarcada a la necesidad de la institución de conocer las causas de la deserción estudiantil.

- El proyecto cuenta con el personal, el investigador Javier José Cevallos Farias, con personal técnico y administrativo del instituto interesados en el tema, con toda la aprobación del Rector y del OCS (Órgano Colégialo Superior), como también docentes y estudiantes.

- La información y datos es recabada desde el sistema SIGALA (sistema integrado de gestión académica Los Andes) que cuenta la institución, como también encuestas a estudiantes, docentes y administrativos. Para el desarrollo de la investigación se utilizó los datos históricos de secretaría desde el año 2019 a 2021, y las encuestas en el periodo académico correspondiente al semestre octubre 2021 marzo 2022.

- Recursos Hardware y Software. - el proyecto utilizo herramientas para la minería de datos de licencia libre y la parte de equipos el investigador asume los costes, dando una factibilidad favorable en cuanto a recursos tecnológicos. Se detalla en la siguiente tabla los recursos que se utilizaron.

Tipo	Herramienta	Pago	libre
Software	Paython		SI
	Googlecolab		SI
	Visual Studio Code		SI
Hardware	portátil Dell core i7	SI	
	Nnternet Netlife	SI	

Tabla 1 Herramientas tecnológicas
Fuente: (Cevallos, 2022)

- ❖ Requisitos, supuestos y restricciones.
  - Requisitos
    - ✓ Autorización de desarrollo del proyecto.
    - ✓ Autorización de las autoridades del Instituto Superior Tecnológico Los Andes.
    - ✓ Disposición de la información necesaria para el proyecto.
    - ✓ Asesoría de expertos en minería de datos.
  - Restricciones
    - ✓ Datos limitados e inválidos

3.3.1.4. Determinación objetivos de la minería de datos

- ❖ Procesar los datos
- ❖ Identificar que variables son las más influyentes en la deserción estudiantil.
- ❖ Predecir considerando otras variables externas para evaluar el comportamiento de los estudiantes durante esos periodos por ejemplo socioeconómicos, pedagógicos, lugar de residencia, etc.
- ❖ Generar un modelo predictivo sobre la deserción estudiantil.
- ❖ Evaluar los modelos.

a) Criterios de éxito de minería de datos.

En este punto se enfoca a la correcta predicción de la deserción estudiantil de la institución educativa, teniendo el menor grado de error.

3.3.1.5. Producción de un plan de proyecto

El plan del proyecto está basado en el cronograma de desarrollo planteado para esta investigación el cual se encuentra en anexos.

3.3.2. Comprensión de los Datos

En esta segunda fase se recolectará, analizará, interpretará y se hará una limpieza de los datos recabados y obtenidos del proyecto.

3.3.2.1. Recolectar datos

Se recolectará datos históricos de estudiantes matriculados, graduados y desertores del año 2019 al 2021. La técnica utilizada para la recolección de datos será la encuesta, que permite tener una perspectiva objetiva y generalizable del problema a describir. En correspondencia con ello, se aplicará el instrumento denominado cuestionario, donde se realizarán diversas preguntas con el objetivo de identificar los causales de la deserción.

Se generará un .xlsx de datos históricos de estudiantes matriculados y desertores con las variables que influyen en la deserción, problemas socioeconómicos, pedagógicos, etc.

La tabla 1 muestra los periodos académicos.

No	Semestre	Periodo
1	Octubre 2018 a marzo 2019	2018-2
2	Abril 2019 a septiembre 2019	2019-1
3	Octubre 2019 a marzo 2020	2019-2
4	Abril 2020 a septiembre 2020	2020-1
5	Octubre 2020 a marzo 2021	2020-2
6	Abril 2021 a septiembre 2021	2021-1
7	Octubre 2021 a marzo 2022	2021-2
8	Abril 2022 a septiembre 2022	2022-1

Tabla 2 Periodos académicos
Fuente: (Istla, 2022)

La Tabla 2 muestra la Población de todos los estudiantes de la institución

Carrera	Estudiantes
Tecnología Superior en desarrollo de software	213
Tecnología Superior en contabilidad	263
Tecnología Superior en electricidad	360
Tecnología Superior en Seguridad Ciudadana y Orden Publico	280
Tecnología Superior en Gestión y Transporte Terrestre	152
Tecnología Superior Diseño de modas	145
Tecnología Superior en actividad física y recreacional	80

Tabla 3 Población
Fuente: (Istla, 2022)

3.3.2.2. Descripción de los datos

Los datos se extraen desde el sistema SIGALA donde constan datos socioeconómicos y académicos de los educandos.

Para la extracción de los datos, se consideró la posibilidad de agrupar otra información de estudiantes que no pertenecía a la fuente propia considerada para la analítica y predicción (se utiliza la tomada de SIGALA). La causa por la que no se utilizaron algunos datos de SIGALA es debido a que no se encuentran cargados ahí, puesto que dicha información y responsabilidad le pertenece a secretaría académica. Por ello, todos los datos que no se han podido sacar de SIGALA, se proponen para ser incluidos al software a través de la ficha de tutoría que tiene el estudiante. Asimismo, se realizará una entrevista a los docentes y coordinadores para obtener información de la parte académica.

La tabla 4 muestra nombres de variables y descripción de las socioeconómicas y académicas.

Campo	Descripción
tipoDocumentoId	donde 1 es cédula, y 2 es pasaporte
numeroIdentificacion	Para cédula: 10 dígitos (ej., 1798630235), Para pasaporte: 9 dígitos, entre números y caracteres alfabéticos en mayúscula (ej., AAE890094)
primerApellido	En este campo ingresar el primer apellido del estudiante matriculado
segundoApellido	En este campo ingresar el segundo apellido del estudiante matriculado
primerNombre	Este campo debe contener el primer nombre de estudiante
segundoNombre	Este campo debe contener el segundo nombre estudiante
sexoId	Donde 1 hombre, 2 mujer
generoId	Donde 1 masculino, 2 femenino
estadocivilId	1 soltero/a, 2 Casado/a, 3 Divorciado/a, 4 Unió libre, 5 Viudo/a
etniaId	1 indígena, 2 Afroecuatoriano, 3 Negro, 4 Mul 5 Montuvio, 6 Mestizo, 7 Blanco, 8 Otro, 9 No registra
pueblonacionalidadId	Esta opción se habilitará únicamente si el estudiante elige la opción 1 (Indígena) del ítem
tipoSangre	1 A+, 2 A-,3 B+, 4 B-,5 AB+, 6 AB-,7 O+, 8 C
discapacidad	donde 1 es SI y 2 es NO
porcentajeDiscapacidad	Si 1, valor entero sin porcentaje (ej.: 70). Si 2, ingresar NA.

numCarnetConadis	Si es 1, son 7 dígitos correspondientes al nro. d carnet (ej., 1729180). Si es2, valor NA
tipoDiscapacidad	1 intelectual, 2 Física, 3 Visual, 4 Auditiva, 5 Mental, 6 Otra, 7 No aplica
fechaNacimiento	formato: aaaa-mm-dd, (ej., 2015-11-25)
paisNacionalidadId	escoger código país
provinciaNacimientoId	escoger código provincia
cantonNacimientoId	escoger código cantón
paisResidenciaId	escoger código país
provinciaResidenciaId	escoger código provincia
cantonResidenciaId	escoger código cantón
tipoColegioId	1 fiscal, 2 Fiscomisional, 3 Particular, 4 Municipal, 5 Extranjero, 6 No registra
modalidadCarrera	1 presencial, 2 Semi-Presencial, 3 Distancia, 4 Dual
jornadaCarrera	1 matutina, 2 Vespertina, 3 Nocturna, 4 Intensi
fechaInicioCarrera	formato: aaaa-mm-dd, (ej., 2015-11-25)
fechaMatrícula	formato: aaaa-mm-dd, (ej., 2015-11-25)
tipoMatriculaId	1 ordinaria, 2 Extraordinaria, 3 Especial
nivelAcademicoQueCursa	1 1ro, 2 2do, 3 3ro, 4 4to, 5 5to, 6 6to
duracionPeriodoAcademico	Número de semanas de duración del periodo académico
haRepetidoAlMenosUnaMateria	donde 1 es SI y 2 es NO
paraleloId	1 A, 2 B, 3 C, 4 D, 5 E, 6 F, 7 G, 8 H, 9 I, 10 J
haPerdidoLaGratuidad	1 Sí, 2 No, 3 No aplica
recibePensionDiferenciada	1 Sí, 2 No, 3 No aplica
studianteocupacionId	1 Solo estudia, 2 Trabaja y estudia
ingresosestudianteId	1 Financiar sus estudios, 2 Para mantener a su hogar, 3 Gastos Personales, 4 No aplica
bonodesarrolloId	donde 1 es SI y 2 es NO
haRealizadoPracticasPreprofesionales	donde 1 es SI y 2 es NO
nroHorasPracticasPreprofesionalesPorPeriodo	Número entero, sin dato es NA
entornoInstitucionalPracticasProfesionales	1 pública, 2 Privada, 3 ONG, 4 Otro, 5 No apli
sectorEconomicoPracticaProfesional	Sector económico en el que realizó las práctica pre profesionales
tipoBecaId	1 total, 2 Parcial, 3 No aplica
primeraRazonBecaId	1 socioeconómica, 2 No aplica
segundaRazonBecaId	1 excelencia Académica, 2 No aplica
terceraRazonBecaId	1 deportista, 2 No aplica
cuartaRazonBecaId	1 pueblos y Nacionalidades, 2 No aplica
quintaRazonBecaId	1 discapacidad, 2 No aplica
sextaRazonBecaId	1 otra 2, No aplica
montoBeca	un número entero, (ej., 30000). Si no NA
porcientoBecaCoberturaArancel	un número entero, (ej., 80). Si no NA
porcientoBecaCoberturaManuntencion	un número entero, (ej., 80). Si no NA
financiamientoBeca	1 fondos propios, 2 Transferencia del Estado, 3 Donaciones, 4 No aplica
montoAyudaEconomica	un número entero, (ej., 3000). Si no NA
montoCreditoEducativo	un número entero, (ej., 3000). Si no NA

participaEnProyectoVinculacionSociedad	1 sí, 2 No, 3 No aplica
tipoAlcanceProyectoVinculacionId	1 nacional, 2 Provincial, 3 Cantonal, 4 Parroqu 5 No aplica
correoElectronico	correo electrónico que la institución asignó al estudiante, SI NO NA
numeroCelular	10 dígitos de dicho número (ej., 0974563897), no (0000000000)
nivelFormacionPadre	1 Centro de Alfabetización, 2 Jardín de infante Primaria, 4 Educación Básica, 5 Secundaria, 6 Educación Media, 7 Superior no Universitaria, Superior Universitaria, 9 Posgrado, 10 No apli
nivelFormacionMadre	1 Centro de Alfabetización, 2 Jardín de infante Primaria, 4 Educación Básica, 5 Secundaria, 6 Educación Media, 7 Superior no Universitaria, Superior Universitaria, 9 Posgrado, 10 No apli
ingresoTotalHogar	número entero (ej., 5000)
cantidadMiembrosHogar	número entero (ej., 6)

Tabla 4 Descripción de los datos socioeconómicas
Fuente: (CACES, 2021)

En la Tabla 5 muestra los nombres de las variables y descripción de las variables académicas.

Nro.	**Nombre**	**Tipo**
1	Costoporsemestre	Float
2	Totalsemestres	Entero
3	Totalmaterias	Entero
4	Materiasporsemestre	Entero
5	Semestresaprobados	Entero
6	Materiasaprobadas	Entero
7	Estadoestudiantetxt	Carácter

Tabla 5 Variables académicas

Fuente: (Istla, 2022)

En la tabla 6 muestra la descripción de cada una de las variables que contiene las columnas de la situación académica.

Campo	**Descripción**
Costoporsemestre	Consta el valor de cada semestre
Totalsemestres	La carrera es de 5 semestres
Totalmaterias	La carrera tiene 30 materias
Materiasporsemestre	Por semestre son 6 materias
Semestresaprobados	Consta los semestres que han aprobado cada estudiante a lo largo de sus estudios
Materiasaprobadas	Consta las materias aprobadas a lo largo de los estudias de cada estudiante
Estadoestudiantetxt	Expresa si el alumno esta graduada, es todavía estudiante o es retirado

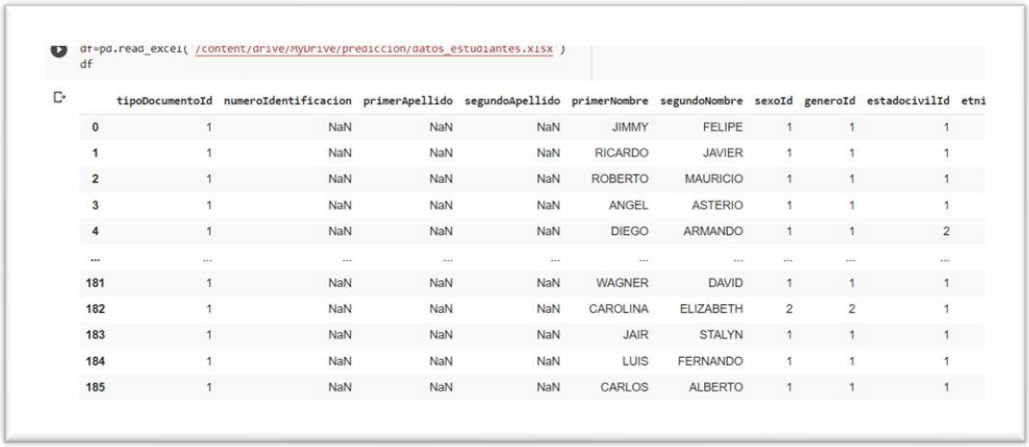
Tabla 6 Descripción de los datos académicos
Fuente: (Istla, 2022)

3.3.2.3. Exploración de los datos

Dando cumplimiento con el primer objetivo específico del proyecto: Determinar las características socioeconómicas de los estudiantes.

Se procede a la recolección de los datos se utilizando el lenguaje de programación Python con todas sus librerías necesarias para realizar un análisis detallado de las variables más relevantes. Con la ayuda del comando pd.read_excel, nos conectamos a nuestra data frame.

En la figura 1 se observa los datos ya leídos desde lenguaje de programación Python.



```
df=pd.read_excel('/content/drive/MyDrive/prediccion/datos_estudiantes.xlsx')
df
```

	tipoDocumentoId	numeroIdentificacion	primerApellido	segundoApellido	primerNombre	segundoNombre	sexoId	generoId	estadocivilId	etni
0	1	NaN	NaN	NaN	JIMMY	FELIPE	1	1	1	
1	1	NaN	NaN	NaN	RICARDO	JAVIER	1	1	1	
2	1	NaN	NaN	NaN	ROBERTO	MAURICIO	1	1	1	
3	1	NaN	NaN	NaN	ANGEL	ASTERIO	1	1	1	
4	1	NaN	NaN	NaN	DIEGO	ARMANDO	1	1	2	
...	...	...	...	...	...	...	...	...	...	
181	1	NaN	NaN	NaN	WAGNER	DAVID	1	1	1	
182	1	NaN	NaN	NaN	CAROLINA	ELIZABETH	2	2	1	
183	1	NaN	NaN	NaN	JAIR	STALYN	1	1	1	
184	1	NaN	NaN	NaN	LUIS	FERNANDO	1	1	1	
185	1	NaN	NaN	NaN	CARLOS	ALBERTO	1	1	1	

Figura 1 Exploración de la dataset
Fuente: (Cevallos, 2022)

Con la ayuda de la librería CountPlot en Python vamos a explorar las variables del proyecto.

En la Figura 2 muestra la distribución por sexo, donde 1 es hombre y 2 es mujer.

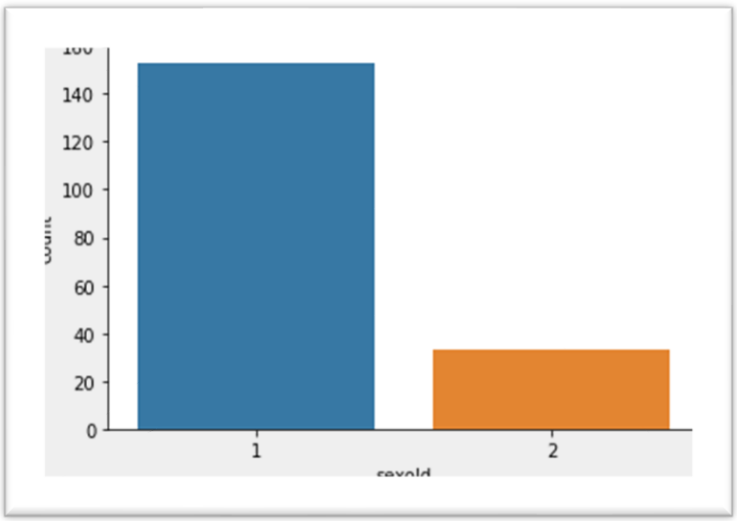


Figura 2 Descripción por género de los estudiantes
Fuente: (Cevallos, 2022)

En la Figura 3 muestra la distribución por género, donde 1 es masculino y 2 es femenino.

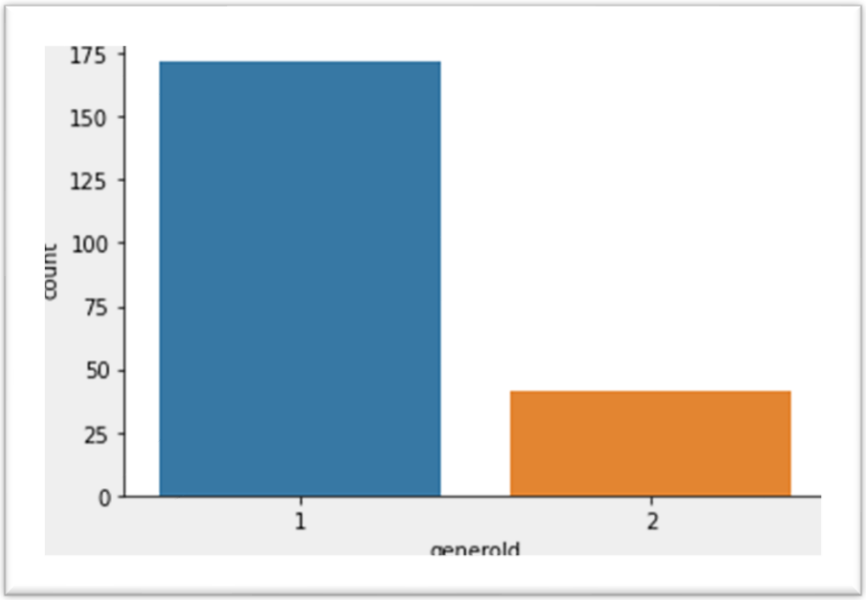


Figura 3 Género
Fuente: (Cevallos, 2022)

En la Figura 4 muestra la distribución de estado civil, donde 1 soltero/a, 2 Casado/a, 3 Divorciado/a, 4 Unión libre, 5 Viudo/a.

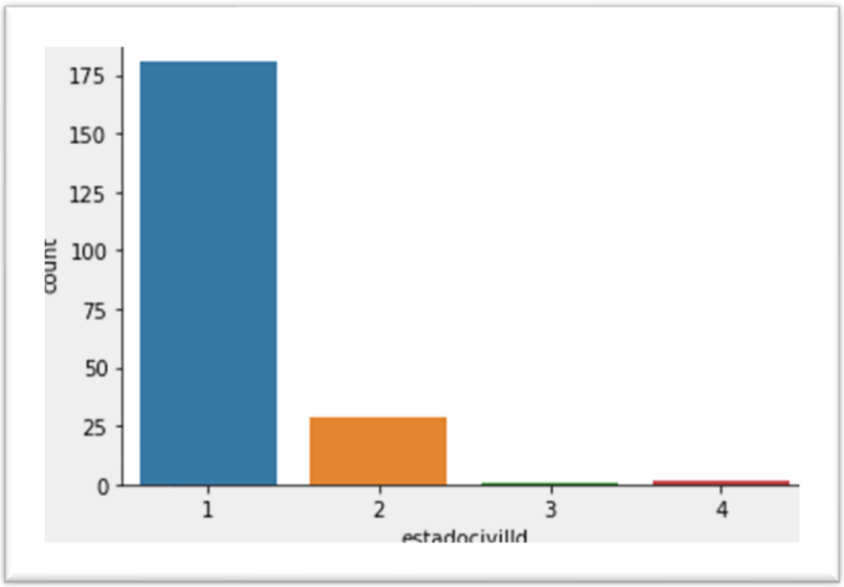


Figura 4 Estado Civil
Fuente: (Cevallos, 2022)

En la Figura 5 muestra la distribución de discapacidad, donde 1 es no y 2 es sí.

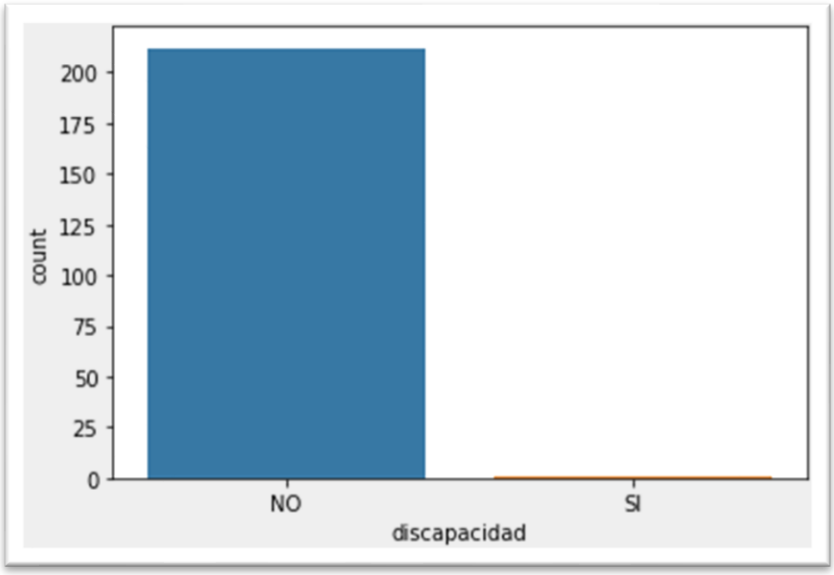


Figura 5 Discapacidad
Fuente: (Cevallos, 2022)

En la Figura 6 muestra la distribución de tipo de colegio, donde 1 fiscal, 2 Fiscomisional, 3 Particular, 4 Municipal, 5 Extranjero, 6 No registra.

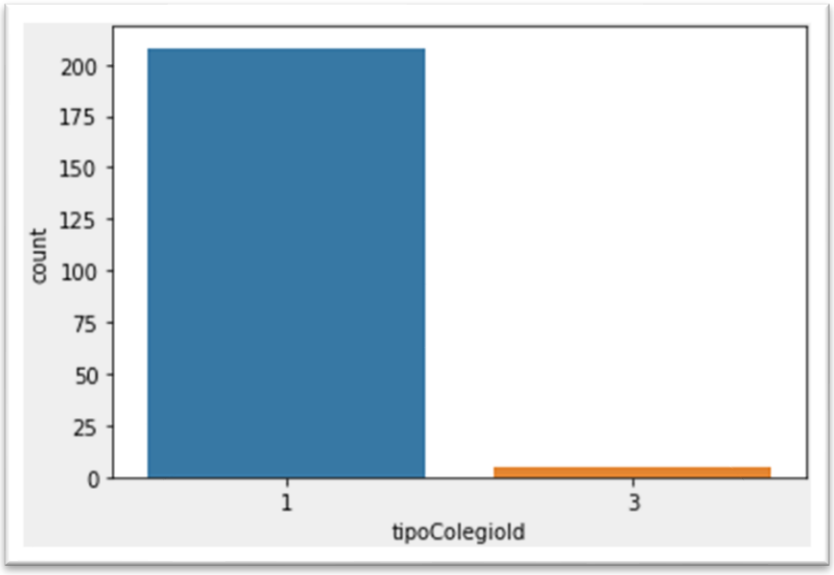


Figura 6 Tipo de colegio
Fuente: (Cevallos, 2022)

En la Figura 7 muestra la distribución de nivel de formación del padre, donde 1 Centro de Alfabetización, 2 Jardín de infantes, 3 Primaria, 4 Educación Básica, 5 Secundaria, 6 Educación Media, 7 Superior no Universitaria, 8 Superior Universitaria, 9 Posgrado, 10 No aplica.

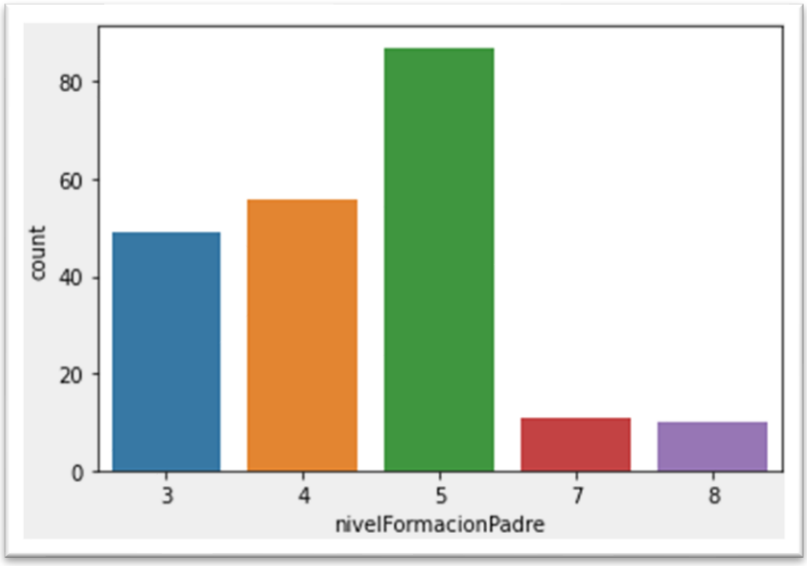


Figura 7 Nivel formación del padre
Fuente: (Cevallos, 2022)

En la Figura 8 muestra la distribución de nivel de formación de la madre, donde 1 Centro de Alfabetización, 2 Jardín de infantes, 3 Primaria, 4 Educación Básica, 5 Secundaria, 6 Educación Media, 7 Superior no Universitaria, 8 Superior Universitaria, 9 Posgrado, 10 No aplica

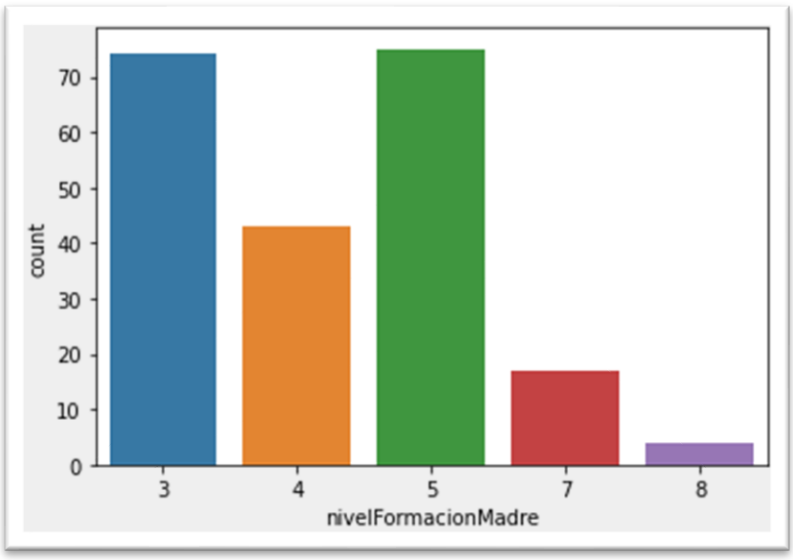


Figura 8 Nivel formación de la madre
Fuente: (Cevallos, 2022)

En la Figura 9 muestra la distribución de ingresos en el hogar.

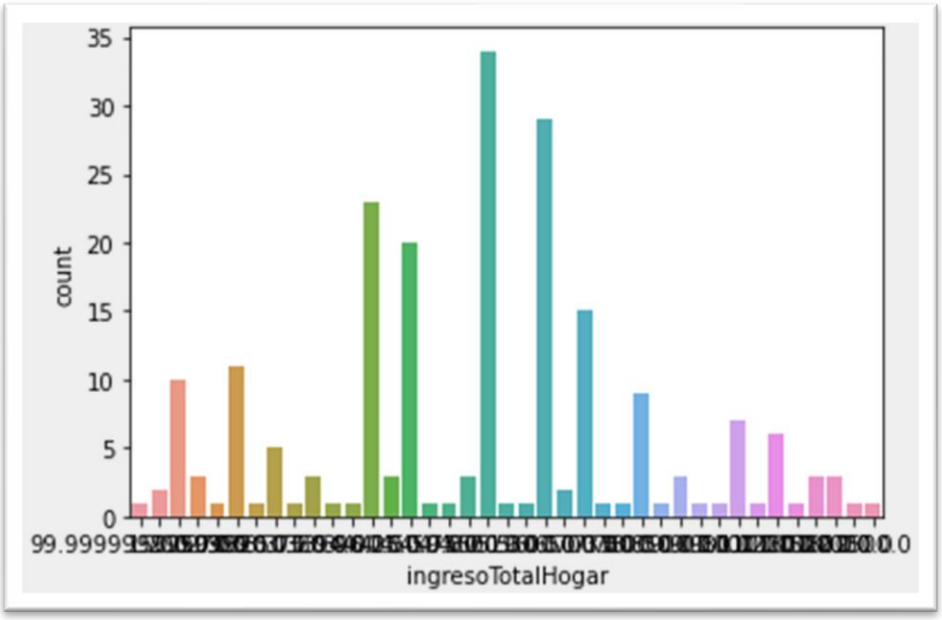


Figura 9 Ingresos en el hogar
Fuente: (Cevallos, 2022)

En la Figura 10 muestra la distribución de cantidad de miembros en el hogar del estudiante

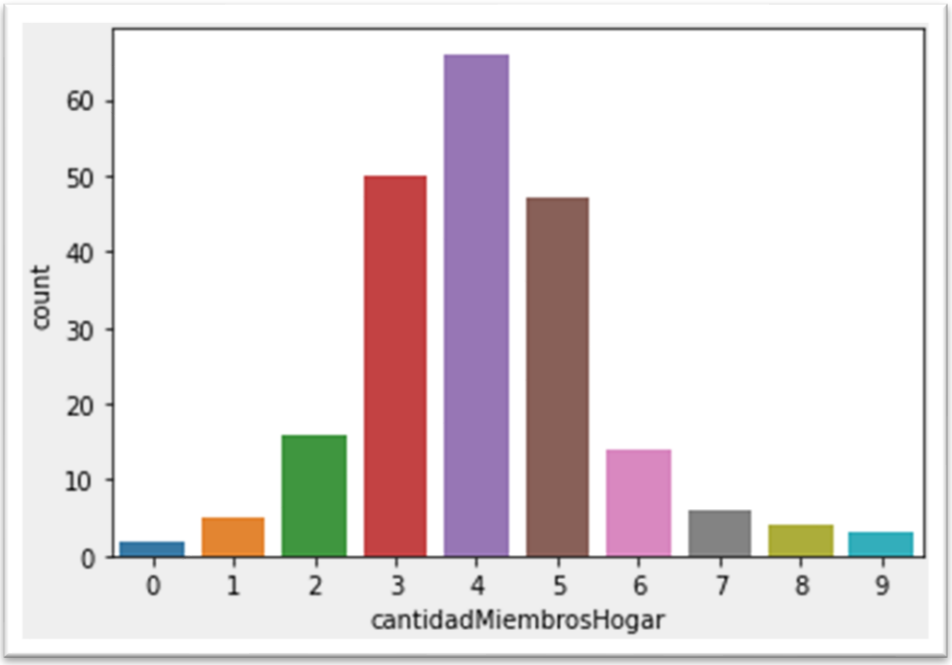


Figura 10 Miembros en el hogar
Fuente: (Cevallos, 2022)

En la Figura 11 muestra la distribución de estado del estudiante, donde 1 está graduado, 2 está estudiando actualmente, 3 está retirado.

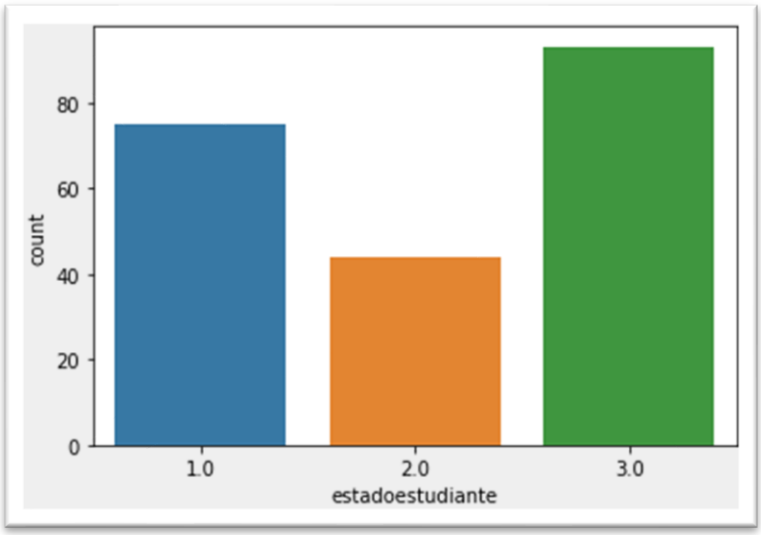


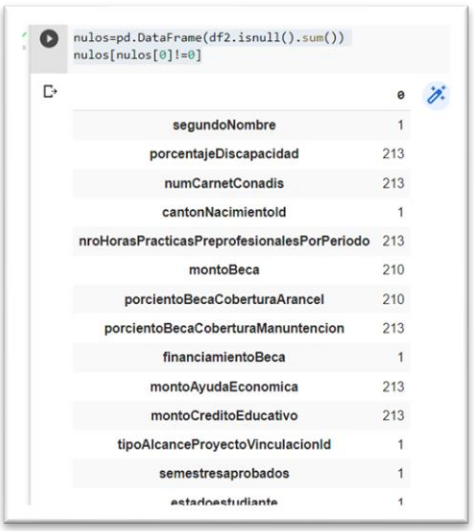
Figura 11 Estado del estudiante
Fuente: (Cevallos, 2022)

3.3.2.4. Verificar la calidad de los datos

En la verificación de datos se utilizó las librerías y comandos de Python, permitiendo obtener datos con características aceptables para su proceso, análisis e interpretación.

Utilizando el comando (isnull) se constató el número de valores nulos por variable.

En la figura 12 da los resultados de número de valores nulos por variable



```
nulos=pd.DataFrame(df2.isnull().sum())
nulos[nulos[0]!=0]
```

	0
segundoNombre	1
porcentajeDiscapacidad	213
numCarnetConadis	213
cantonNacimientoId	1
nroHorasPracticasPreprofesionalesPorPeriodo	213
montoBeca	210
porcientoBecaCoberturaArancel	210
porcientoBecaCoberturaManuntencion	213
financiamientoBeca	1
montoAyudaEconomica	213
montoCreditoEducativo	213
tipoAlcanceProyectoVinculacionId	1
semestresaprobados	1
estadoestudiante	1

Figura 12 Valores nulos
Fuente: (Cevallos, 2022)

La Figura 13 refleja la proporción de los valores nulos por variable

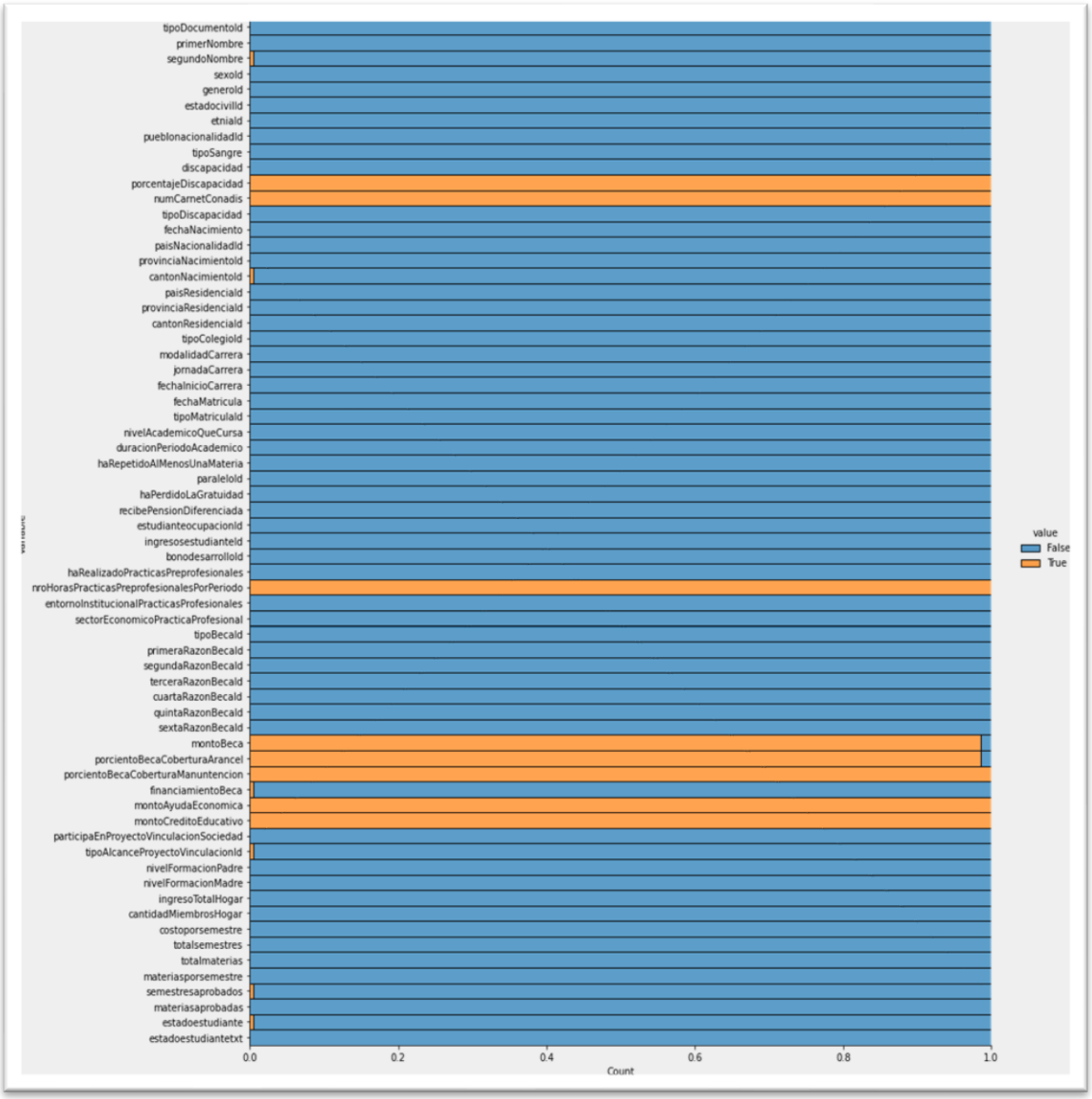


Figura 13 Proporción de los valores nulos
Fuente: (Cevallos, 2022)

Una vez que se realizó la exploración de los datos se pudo constatar que se debía realizar alugas operaciones:

- Eliminar columnas de datos críticos de cada estudiante, con la finalidad de resguardar la integridad de cada uno d ellos mismos, esto apegado a la ética profesional, las columnas que se eliminaron fueron: 'numeroIdentificacion', 'primerApellido', 'segundoApellido', 'correoElectronico', 'numeroCelular

- Se utilizo el comando drop y se cero otra data frame (df2)
- Se reviso que la información que los datos sean correctos y relevantes.
- Se elimina columnas que después de su descripción no contienen información relevante para el estudio estas columnas son:
  - 'tipoDocumentoId'
  - 'primerNombre'
  - 'segundoNombre'
  - 'segundoNombre'
  - 'generoId'
  - 'pueblonacionalidadId'
  - 'tipoSangre'
  - 'porcentajeDiscapacidad'
  - 'numCarnetConadis'
  - 'tipoDiscapacidad'
  - 'paisNacionalidadId'
  - 'provinciaNacimientoId'
  - 'cantonNacimientoId',
  - 'paraleloId'
  - 'bonodesarrolloId'
  - 'nroHorasPracticasPreprofesionalesPorPeriodo'
  - 'entornoInstitucionalPracticasProfesionales'
  - 'sectorEconomicoPracticaProfesional'
  - 'primeraRazonBecaId','segundaRazonBecaId'
  - 'terceraRazonBecaId'
  - 'cuartaRazonBecaId',

- 'quintaRazonBecaId'
- 'sextaRazonBecaId'
- 'montoBeca',
- 'porcientoBecaCoberturaArancel'
- 'porcientoBecaCoberturaManuntencion'
- 'montoAyudaEconomica'
- 'montoCreditoEducativo'
- 'tipoAlcanceProyectoVinculacionId'
- 'haRealizadoPracticasPreprofesionales'
- 'tipoBecaId'
- 'financiamientoBeca'
- 'participaEnProyectoVinculacionSociedad'
- 'tipoMatriculaId'
- 'nivelAcademicoQueCursa'
- 'duracionPeriodoAcademico'
- 'recibePensionDiferenciada'
- 'haPerdidoLaGratuidad'

❖ Generamos una nueva data frame (df3)

3.3.3. Preparación de los Datos

Dando cumplimiento al segundo objetivo del proyecto: Aplicar métricas estadísticas para analizar las variables que ocasionan la deserción estudiantil.

3.3.3.1. Selección de datos

La data contiene información histórica de todas las carreras, para el presente estudio se va utilizar los datos de la carrera Desarrollo de software (sistemas), esto se lo va a realizar por medio del lenguaje de programación Python y sus librerías.

A medida que avance el proyecto se vio la necesidad de seleccionar diferentes tipos de datos el cual contiene la data llamada datos_estudiantes.xlsx.

Selección de atributos. Los datos a seleccionar tendrán información relevante y confidencial de los estudiantes durante el proyecto, pudiendo filtrar socioeconómicos, académicos.

En la Figura 14 observamos ya las columnas con datos relevantes para la realización de proyecto de deserción estudiantil

	sexoid	estadocivilid	etniaid	discapacidad	fechanacimiento	paisresidencia id	provinciaresidencia id	cantonresidencia id	tipocolegioid
0	1	1	6	NO	1995-07-14	56	23	2301	1
1	1	1	6	NO	1990-09-29	56	23	2301	1
2	1	1	6	NO	1994-05-13	56	23	2301	1
3	1	1	6	NO	2000-03-13	56	23	2301	1
4	1	2	6	NO	1982-02-16	56	23	2301	1
...	...	...	...	...	...	...	...	...	...
208	2	1	6	NO	2004-10-27	56	23	2301	1
209	2	1	6	NO	1991-12-31	56	23	2301	1
210	1	1	6	NO	2001-12-11	56	23	2301	1
211	1	1	6	NO	1999-12-25	56	23	2301	1
212	2	1	6	NO	1985-06-04	56	23	2301	1

213 rows x 27 columns

Figura 14 Selección de datos
Fuente: (Cevallos, 2022)

3.3.3.2. Limpieza de datos

El estudio en cuestión utilizo datos en formato xlsx, los mismos que fueron facilitados desde el departamento de secretaria previo autorización de las autoridades. Se deja constancia que los datos facilitados por el departamento encargado son los que la institución carga al Snescyt y que en su mayoría ya estaban verificado, y con la utilización de las diferentes librerías y herramientas del lenguaje de programación Python donde se plasmó la limpieza de las diferentes columnas:

- Se realizó un isnull().sum(), para verificar datos nulos

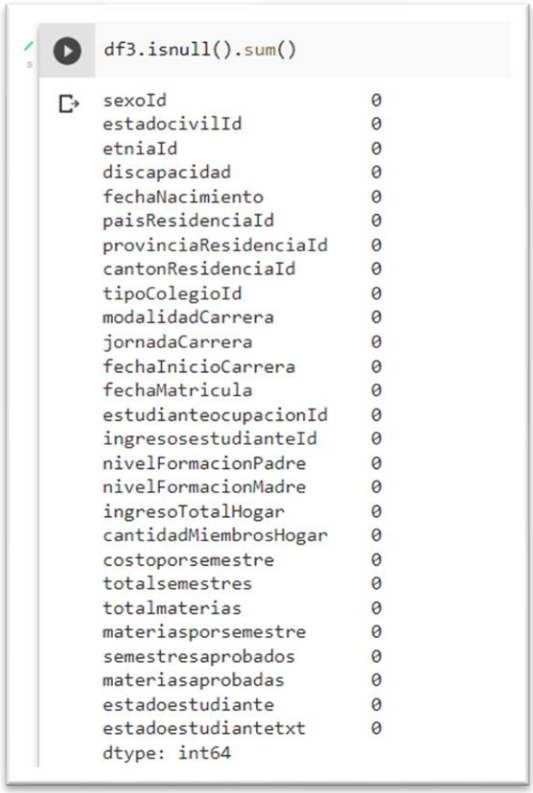
En la Figura 15 muestra los valores nulos, en este caso las variables que tienen nulos son semestresaprobados y estadoestudiante.

```
df3.isnull().sum()

sexoId                  0
estadocivilId           0
etniaId                 0
discapacidad            0
fechaNacimiento         0
paisResidenciaId        0
provinciaResidenciaId   0
cantonResidenciaId      0
tipoColegioId           0
modalidadCarrera        0
jornadaCarrera          0
fechaInicioCarrera      0
fechaMatricula          0
estudianteocupacionId   0
ingresosestudianteId    0
nivelFormacionPadre     0
nivelFormacionMadre     0
ingresoTotalHogar       0
cantidadMiembrosHogar   0
costoporsemestre        0
totalsemestres          0
totalmaterias           0
materiasporsemestre     0
semestresaprobados      1
materiasaprobadas       0
estadoestudiante        1
estadoestudiantetxt     0
dtype: int64
```

Figura 15 Verificar datos nulos
Fuente: (Cevallos, 2022)

- Luego de ver el reporte se procedió a llenar los na de objetos con n/a, utilizando los siguientes comandos
  - df3['semestresaprobados'].fillna(3,inplace=True)
  - df3['estadoestudiante'].fillna(3,inplace=True)
- Se realizó un (df3.isnull().sum().sum()), constatando que no existen valores nulos

En la Figura 16 observamos que ya no existen valores nulos.

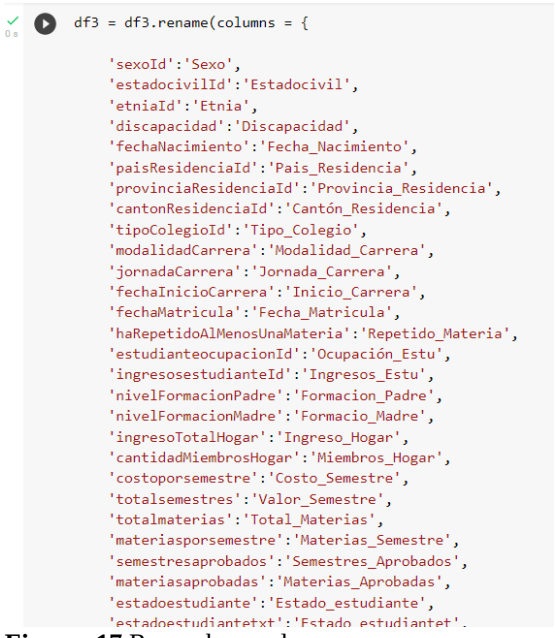


```
df3.isnull().sum()

sexoId                    0
estadocivilId             0
etniaId                   0
discapacidad              0
fechaNacimiento           0
paisResidenciaId          0
provinciaResidenciaId     0
cantonResidenciaId        0
tipoColegioId             0
modalidadCarrera          0
jornadaCarrera            0
fechaInicioCarrera        0
fechaMatricula            0
estudianteocupacionId     0
ingresosestudianteId      0
nivelFormacionPadre       0
nivelFormacionMadre       0
ingresoTotalHogar         0
cantidadMiembrosHogar     0
costoporsemestre          0
totalsemestres            0
totalmaterias             0
materiasporsemestre       0
semestresaprobados        0
materiasaprobadas         0
estadoestudiante          0
estadoestudiantetxt       0
dtype: int64
```

Figura 16 Existencia de datos nulos
Fuente: (Cevallos, 2022)

- Renombramos las columnas

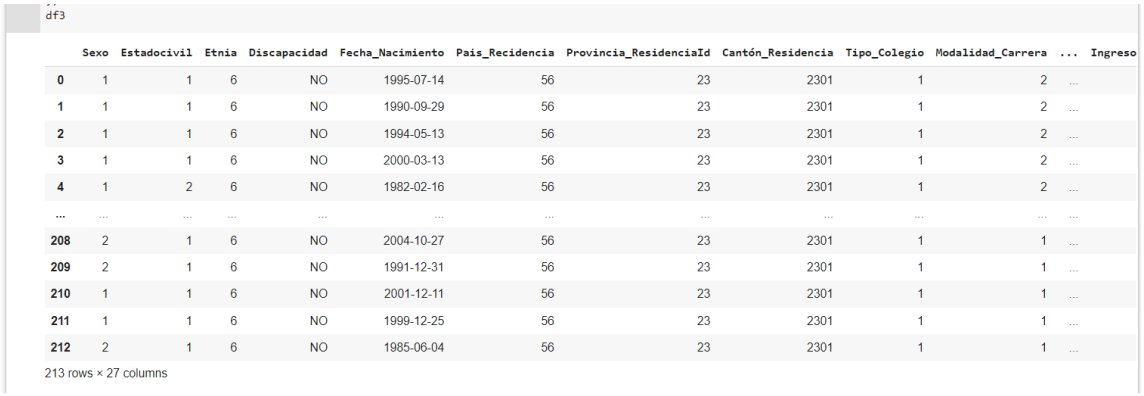
En la Figura 17 se observa los nombres que van a tener las columnas al ser renombradas.



```
df3 = df3.rename(columns = {

    'sexoId':'Sexo',
    'estadocivilId':'Estadocivil',
    'etniaId':'Etnia',
    'discapacidad':'Discapacidad',
    'fechaNacimiento':'Fecha_Nacimiento',
    'paisResidenciaId':'Pais_Residencia',
    'provinciaResidenciaId':'Provincia_Residencia',
    'cantonResidenciaId':'Cantón_Residencia',
    'tipoColegioId':'Tipo_Colegio',
    'modalidadCarrera':'Modalidad_Carrera',
    'jornadaCarrera':'Jornada_Carrera',
    'fechaInicioCarrera':'Inicio_Carrera',
    'fechaMatricula':'Fecha_Matricula',
    'haRepetidoAlMenosUnaMateria':'Repetido_Materia',
    'estudianteocupacionId':'Ocupación_Estu',
    'ingresosestudianteId':'Ingresos_Estu',
    'nivelFormacionPadre':'Formacion_Padre',
    'nivelFormacionMadre':'Formacio_Madre',
    'ingresoTotalHogar':'Ingreso_Hogar',
    'cantidadMiembrosHogar':'Miembros_Hogar',
    'costoporsemestre':'Costo_Semestre',
    'totalsemestres':'Valor_Semestre',
    'totalmaterias':'Total_Materias',
    'materiasporsemestre':'Materias_Semestre',
    'semestresaprobados':'Semestres_Aprobados',
    'materiasaprobadas':'Materias_Aprobadas',
    'estadoestudiante':'Estado_estudiante',
    'estadoestudiantetxt':'Estado_estudiantet',
```

Figura 17 Renombrar columnas
Fuente: (Cevallos, 2022)

En la Figura 18 se visualiza ya los nuevos nombres en la data frame.



df3

	Sexo	Estadocivil	Etnia	Discapacidad	Fecha_Nacimiento	Pais_Recidencia	Provincia_ResidenciaId	Cantón_Residencia	Tipo_Colegio	Modalidad_Carrera	...	Ingreso
0	1	1	6	NO	1995-07-14	56	23	2301	1	2	...	
1	1	1	6	NO	1990-09-29	56	23	2301	1	2	...	
2	1	1	6	NO	1994-05-13	56	23	2301	1	2	...	
3	1	1	6	NO	2000-03-13	56	23	2301	1	2	...	
4	1	2	6	NO	1982-02-16	56	23	2301	1	2	...	
...	...	...	...	...	...	...	...	...	...	...	...	...
208	2	1	6	NO	2004-10-27	56	23	2301	1	1	...	
209	2	1	6	NO	1991-12-31	56	23	2301	1	1	...	
210	1	1	6	NO	2001-12-11	56	23	2301	1	1	...	
211	1	1	6	NO	1999-12-25	56	23	2301	1	1	...	
212	2	1	6	NO	1985-06-04	56	23	2301	1	1	...	

213 rows × 27 columns

Figura 18 Nuevos nombres en la data frame
Fuente: (Cevallos, 2022)

- Se analizo con este comando (df3.dtypes), los tipos de variables y si están acorde a los datos.

En la Figura 19 se visualiza que las fechas de nacimiento, de inicio de carrera y de matrícula están tipo de variable incorrecto. Así mismo costo por semestre debe estar en float, y estado de estudiante debe estar en int

```
df3.dtypes
```

```
sexoId                    int64
estadocivilId             int64
etniaId                   int64
discapacidad             object
fechaNacimiento          object
paisResidenciaId          int64
provinciaResidenciaId     int64
cantonResidenciaId        int64
tipoColegioId             int64
modalidadCarrera          int64
jornadaCarrera            int64
fechaInicioCarrera       object
fechaMatricula           object
estudianteocupacionId     int64
ingresosestudianteId      int64
nivelFormacionPadre       int64
nivelFormacionMadre       int64
ingresoTotalHogar       float64
cantidadMiembrosHogar     int64
costoporsemestre          int64
totalsemestres            int64
totalmaterias             int64
materiasporsemestre       int64
semestresaprobados      float64
materiasaprobadas         int64
estadoestudiante        float64
estadoestudiantetxt      object
dtype: object
```

Figura 19 Tipos de variable
Fuente: (Cevallos, 2022)

- Se realiza cambio del formato de la columna fecha y se adiciona Año, mes y día.

Se procedió con la ayuda del comando (pd.to_datetime), se procedió a cambiar a todas las comunas que contenían fechas, nacimiento, inicio de carrera y fecha de matrícula, con la siguiente línea de código

df3['Fecha_Nacimiento']=pd.to_datetime(df3['Fecha_Nacimiento']).

- Volvemos a describir (df3.dtypes), para corroborar si esta cambiado

La figura 20 en contraste a la figura 19, se puede verificar que los tipos de datos de las variables se han modificado.

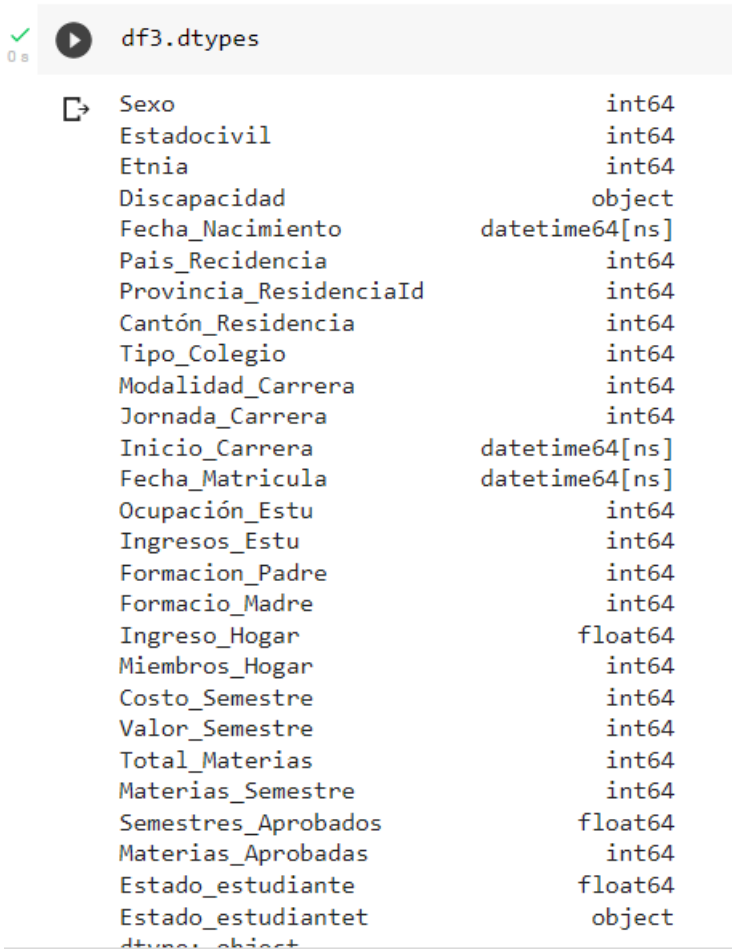


Figura 20 Revisión de fechas
Fuente: (Cevallos, 2022)

Una vez realizado todo el proceso de curación de datos el total de discentes es 213 y un total de columnas 31, los cuales se los analizo descriptivamente de la siguiente manera:

Se procede a realizar un nuevo análisis de datos con la nueva data, utilizando las herramientas (sweetviz, pandas_profiling), ayudó a clasificar las variables categóricas y numéricas con datos estadísticos importantes, con la finalidad de obtener variables relevantes para realizar una matriz de correlación adecuada.

- ❖ Graficas analíticas descriptivas de las variables categórica

La figura 21 muestra que el 81% son Hombres y el 19% son mujeres.

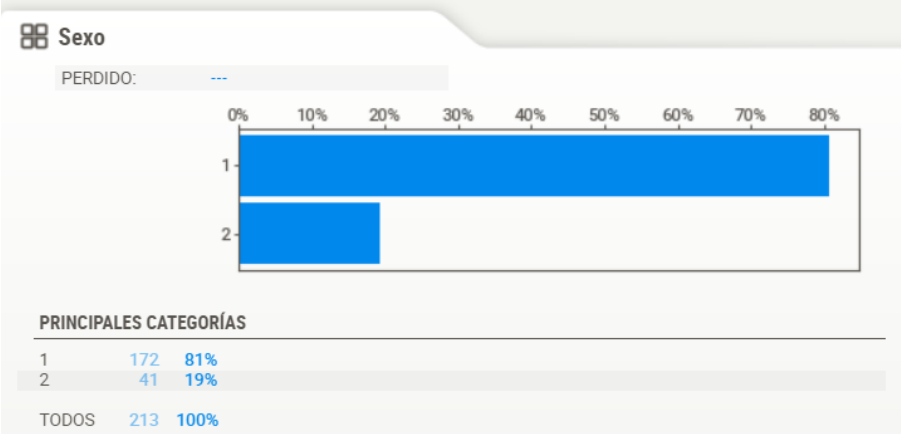


Figura 21 Datos descriptivos de la variable Sexo
Fuente: (Cevallos, 2022)

La figura 22 que el 85% son solteros, un 14% casados, el 2% unión libre y menos del 1% viudo

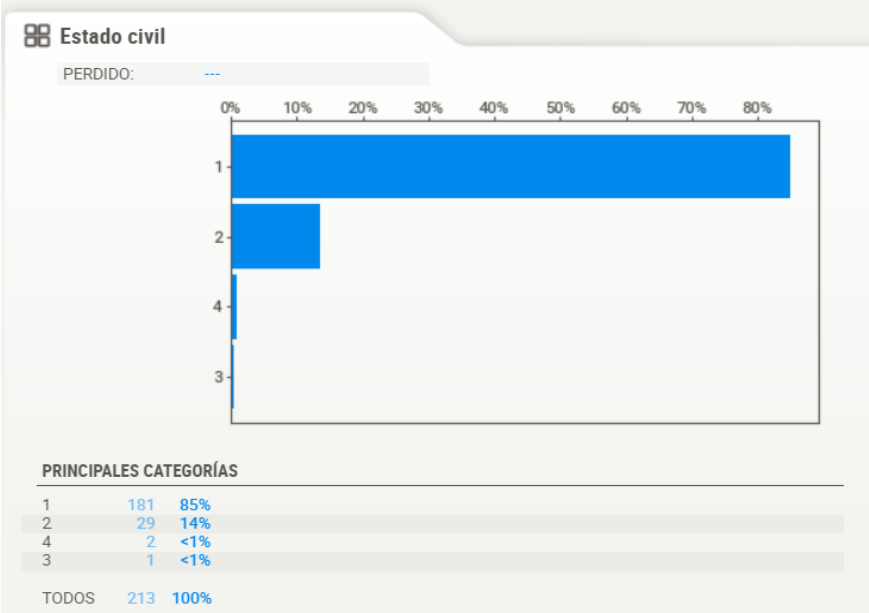


Figura 22 Datos descriptivos de la variable Estado Civil
Fuente: (Cevallos, 2022)

La Figura 23 muestra que el 99% son mestizos, y el 1% negros

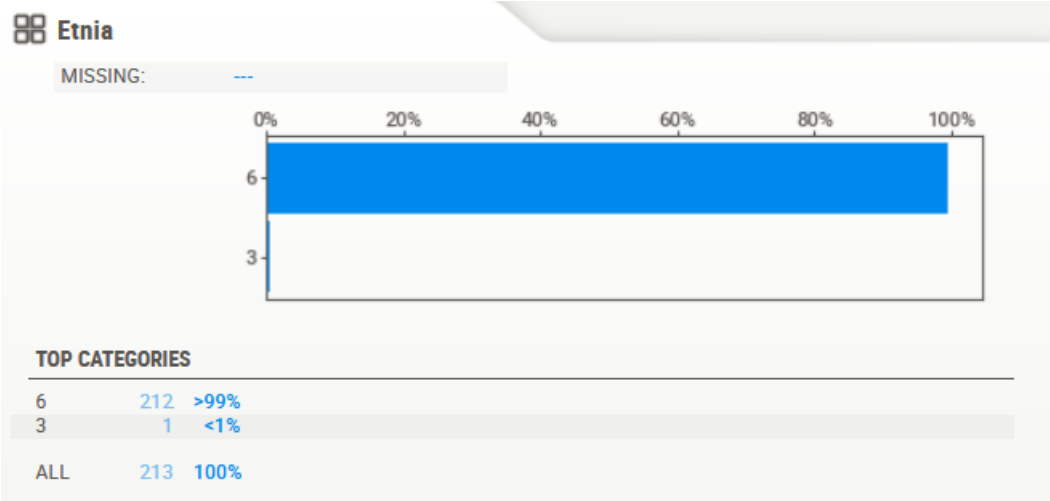


Figura 23 Datos descriptivos de la variable Etnia
Fuente: (Cevallos, 2022)

En la Figura 24 muestra que el 99% de estudiantes pertenecen a la provincia Tsáchilas, y 1% a otras.

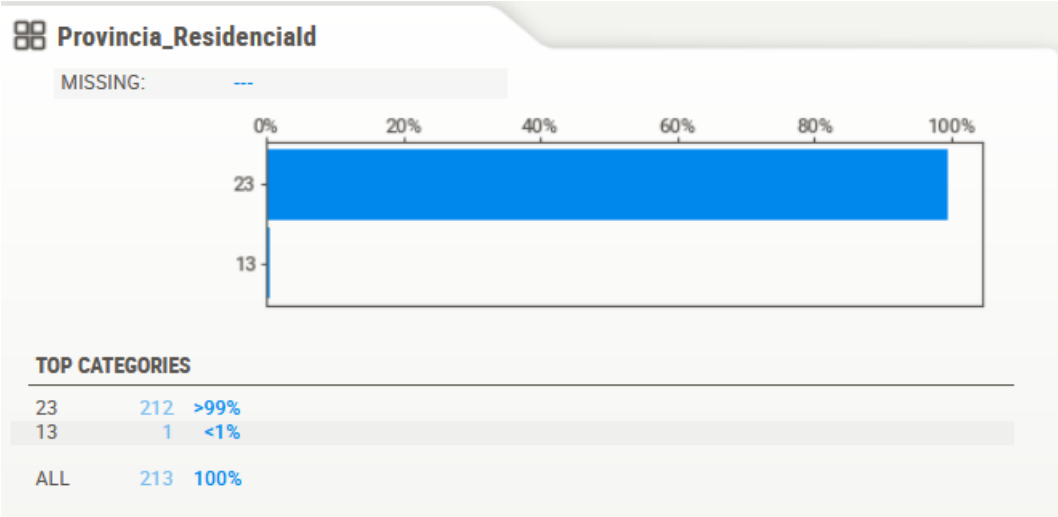


Figura 24 Datos descriptivos de la variable Provincia Residencia
Fuente: (Cevallos, 2022)

La figura 25 muestra que la mayoría de los estudiantes son del Cantón Santo Domingo con el 99%.

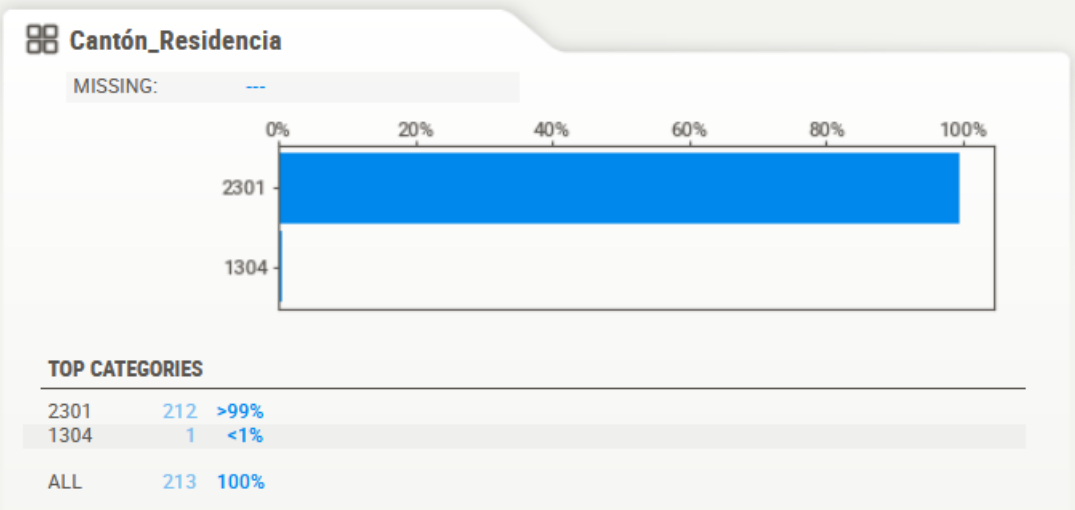


Figura 25 Datos descriptivos de la variable Estado Civil
Fuente: (Cevallos, 2022)

La Figura 26 muestra que el 98% de estudiantes proceden de un colegio fiscal, ante el 2% de un particular.

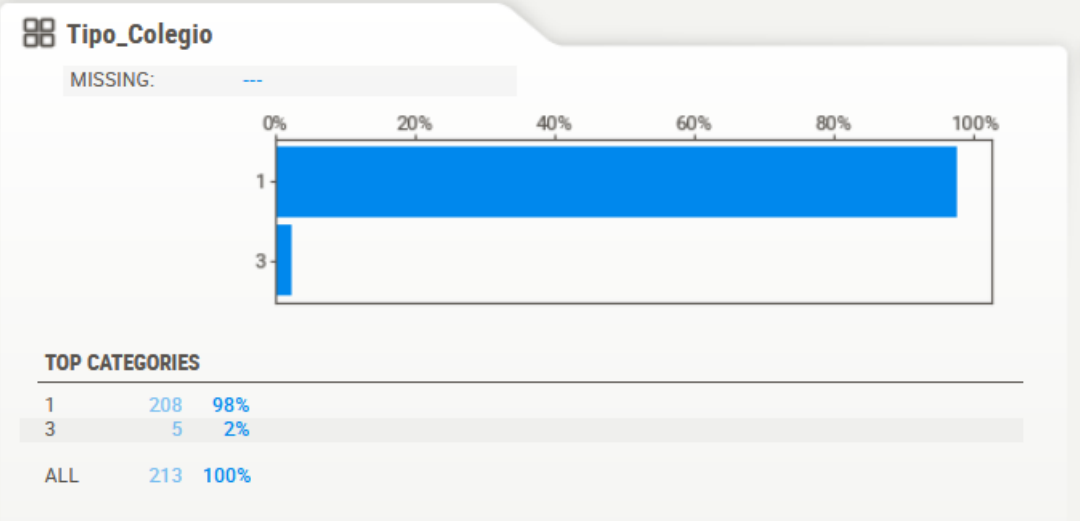


Figura 26 Datos descriptivos de la variable Estado Civil
Fuente: (Cevallos, 2022)

En la figura 27 muestra que un 90 % estudia presencial, ante el 9% semipresencial.

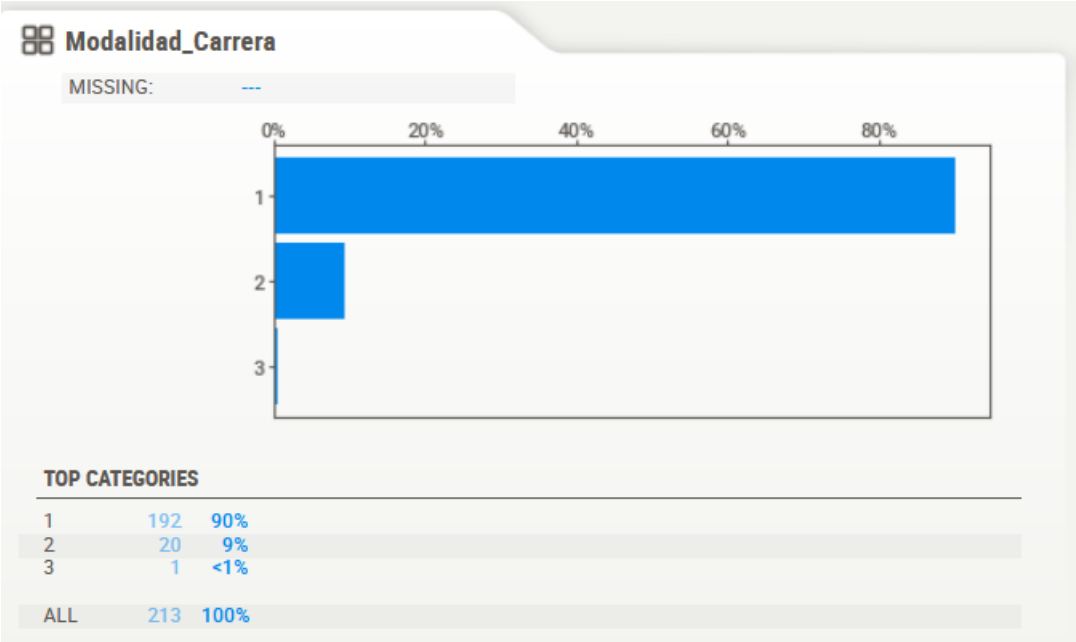


Figura 27 Datos descriptivos de la variable Modalidad Carrera
Fuente: (Cevallos, 2022)

La figura 28 muestra que el 82% utilizan sus ingresos para financiar sus estudios. El 17 % no aplica, el 1 % para mantear su hogar

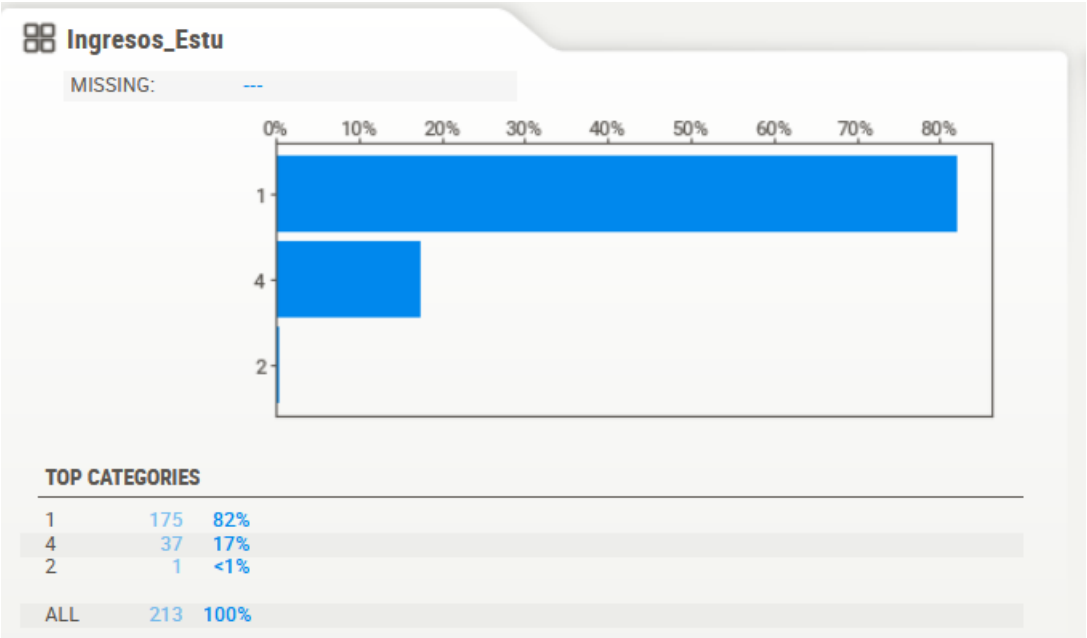


Figura 28 Datos descriptivos de la variable Ingresos Estudiantes
Fuente: (Cevallos, 2022)

La figura 29 muestra que el 41% ha estudiado la secundaria, el 26% educación básica, el 23% primaria, el 5% Superior no Universitaria, un 5% Superior Universitaria.

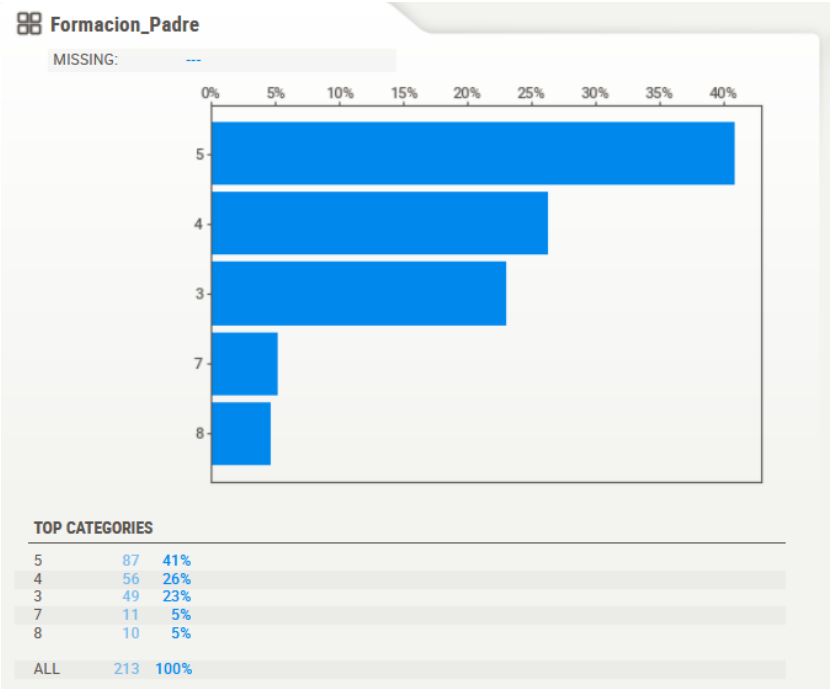


Figura 29 Datos descriptivos de la variable Formación del Padre
Fuente: (Cevallos, 2022)

La figura 30 muestra que el 35% ha estudiado la secundaria, el 35% primaria, el 20% Educación Básica, el 8% superior no universitaria, un 2% Superior Universitaria.

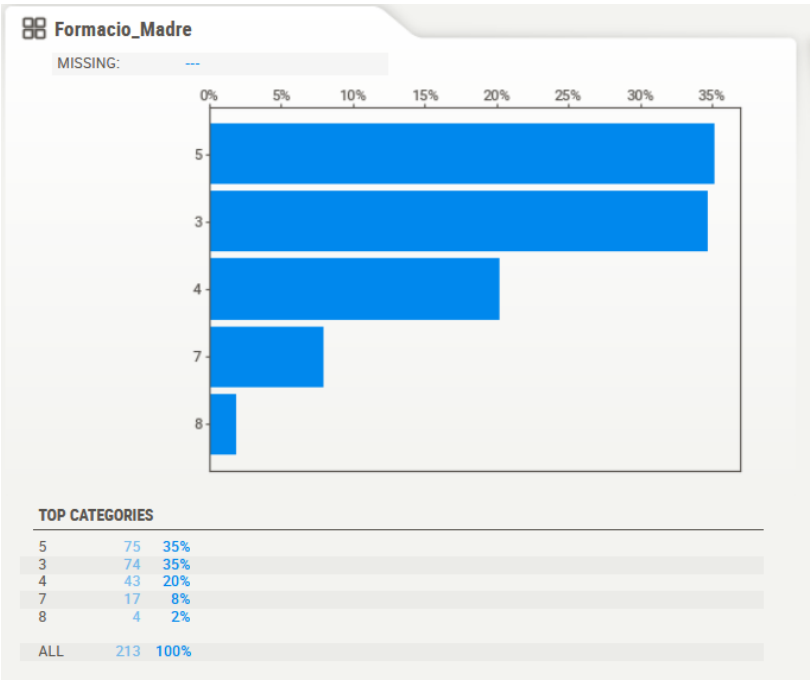


Figura 30 Datos descriptivos de la variable Formación de la Madre
Fuente: (Cevallos, 2022)

La Figura 31 muestra que el 41% han aprobado 5 semestres, el 33% 1 semestre, el 13% 2 semestres, el 8% 4 semestres, el 4% 3 semestres, menos del 1% ningún semestre.

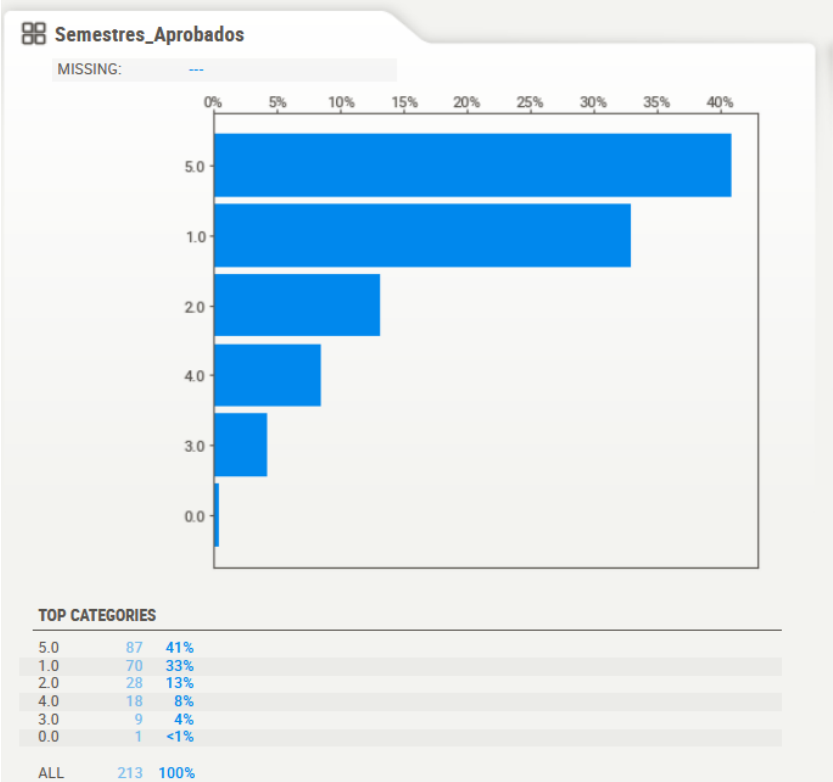


Figura 31 Datos descriptivos de la variable Semestres Aprobados
Fuente: (Cevallos, 2022)

La Figura 32 muestra que el 44% de estudiantes se han retirado a lo largo del tiempo, el 35% se ha graduado mientras que el 21% está estudiando actualmente.

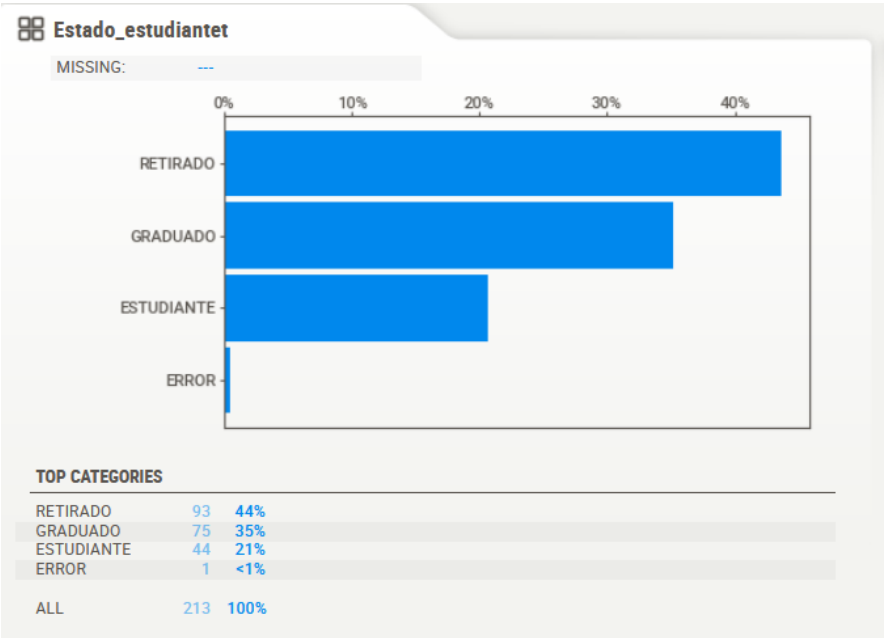


Figura 32 Datos descriptivos de la variable Estado del estudiante
Fuente: (Cevallos, 2022)

❖ Variables Numéricas

La Figura 33 muestra que esta variable numérica Ingreso en el hogar, está asociada a otras categorías como Fecha_Matricula, Formacion_Padre, Formacio_Madre, Miembros_Hogar, Inicio_Carrera, Ingresos_Estu, Ocupación_Estu, Año, Semestres_Aprobados, Estadocivil, Estado_estudiante, Estado_estudiantet, Modalidad_Carrera

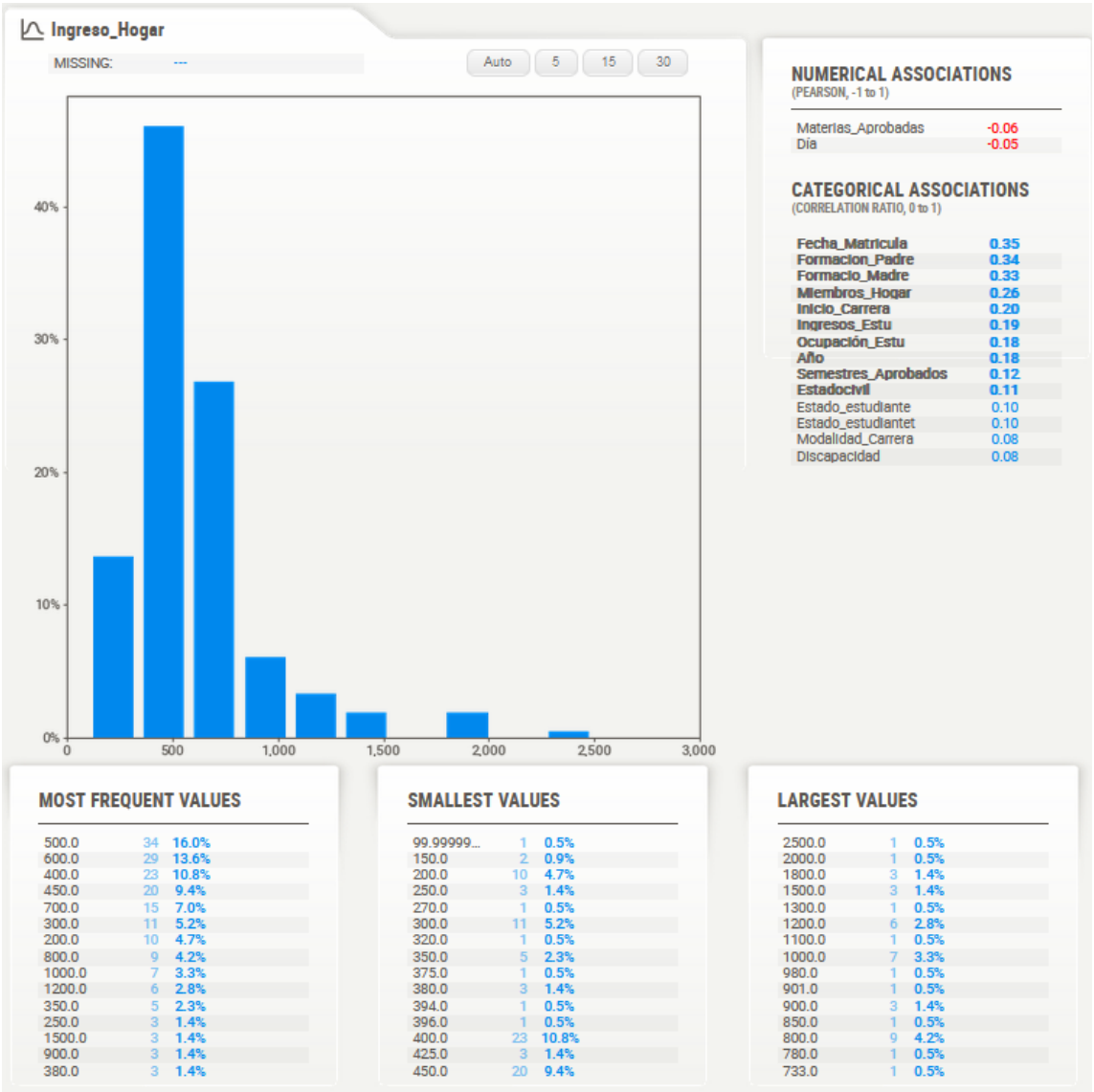


Figura 33 Datos descriptivos de la variable Ingreso Hogar
Fuente: (Cevallos, 2022)

La Figura 34 muestra que la variable Miembros en el Hogar tiene relación de Ingreso ene l Hogar, Materias Aprobadas.

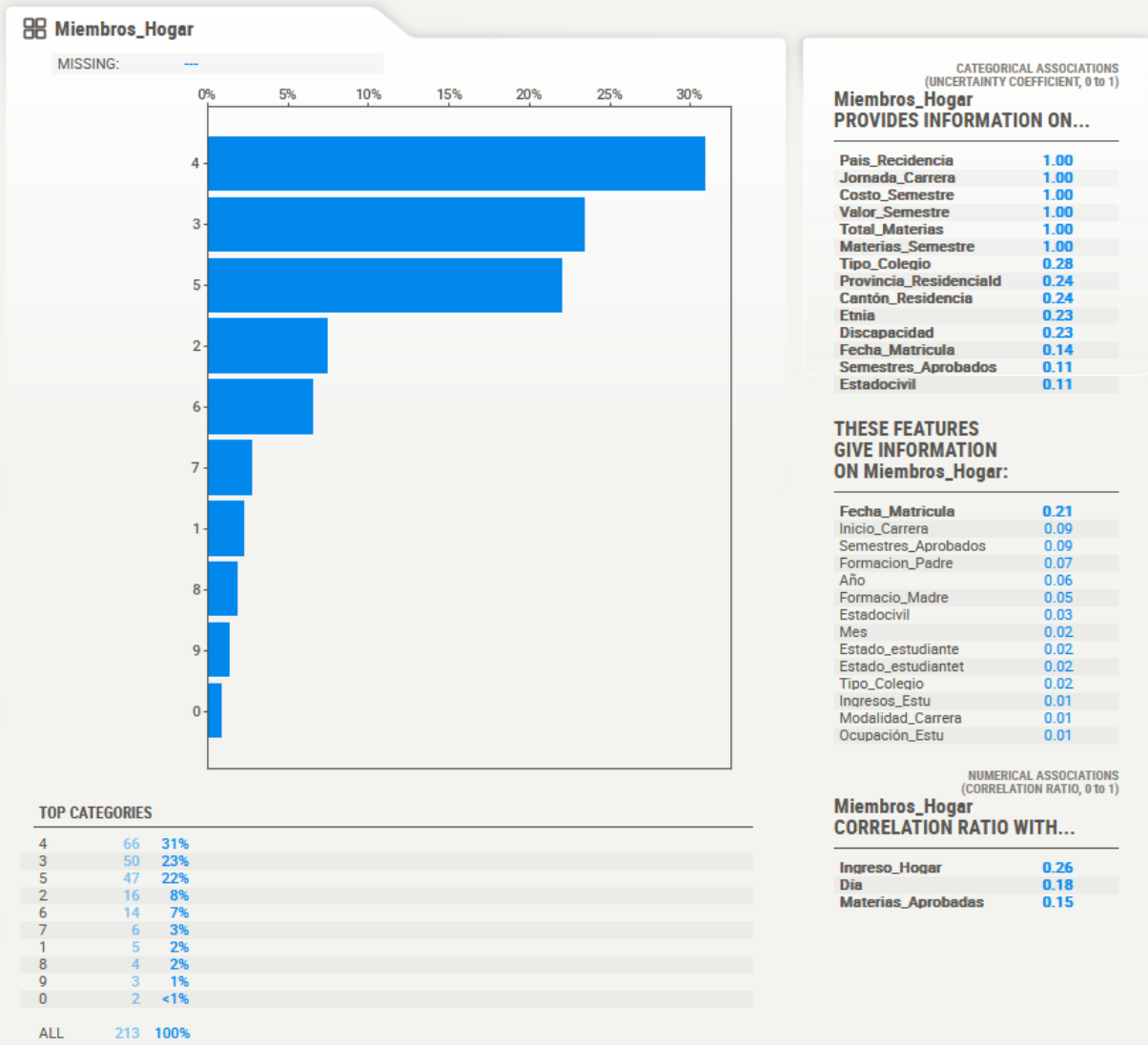


Figura 34 Datos descriptivos de la variable Miembros del Hogar
Fuente: (Cevallos, 2022)

Figura 35 muestra que esta variable costo de los semestres es constante y no tienen relación relevante con otras categorías.

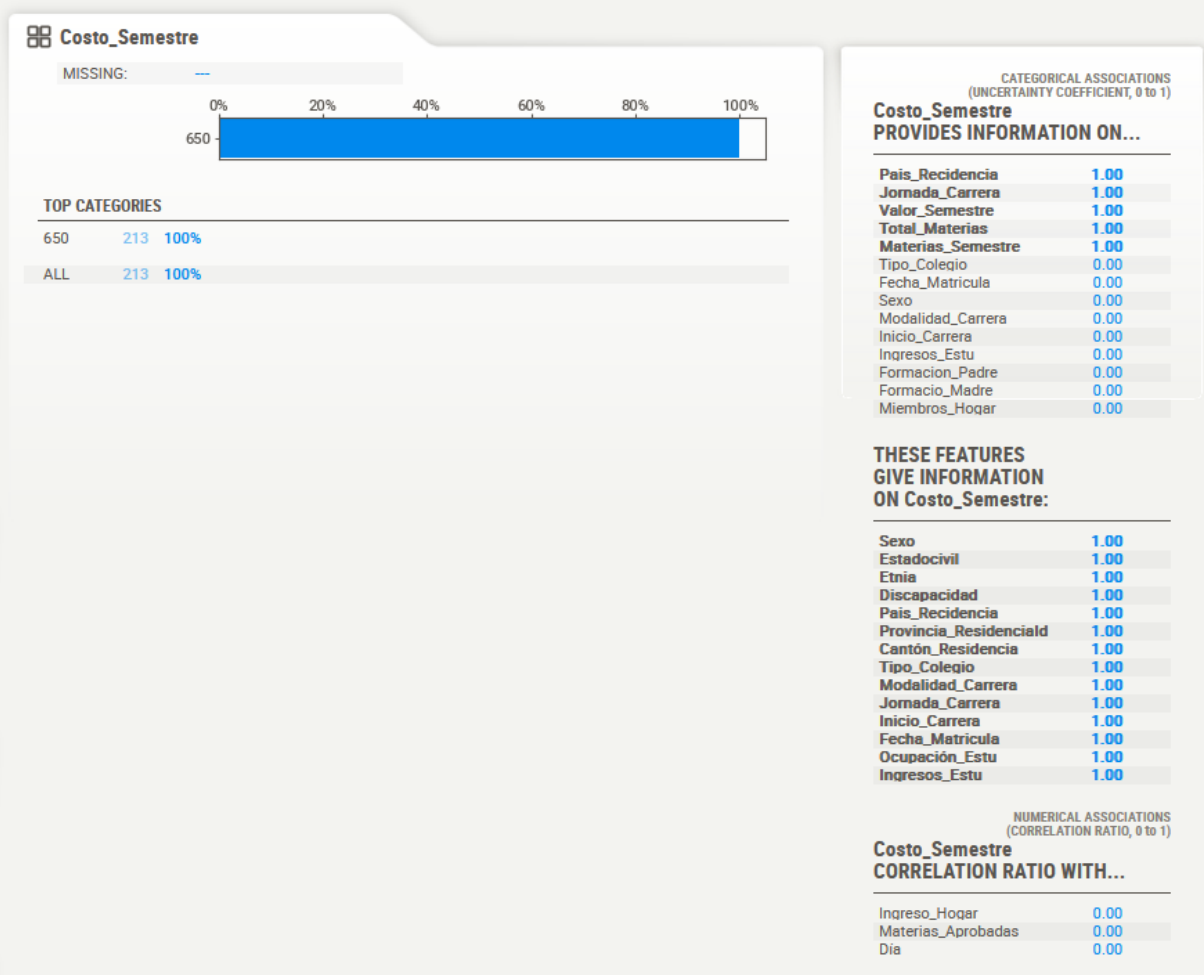


Figura 35 Datos descriptivos de la variable Costo de cada Semestre
Fuente: (Cevallos, 2022)

La figura 36 nos muestra que la variable Materias Aprobadas esta relacionada con otras categorías como, Semestres_Aprobados, Estado_estudiante, Estado_estudiantet, Fecha_Matricula, Inicio_Carrera, Año, Formacio_Madre, Formacion_Padre, Mes, Ingresos_Estu, Ocupación_Estu, Miembros_Hogar, Estadocivi, Discapacidad

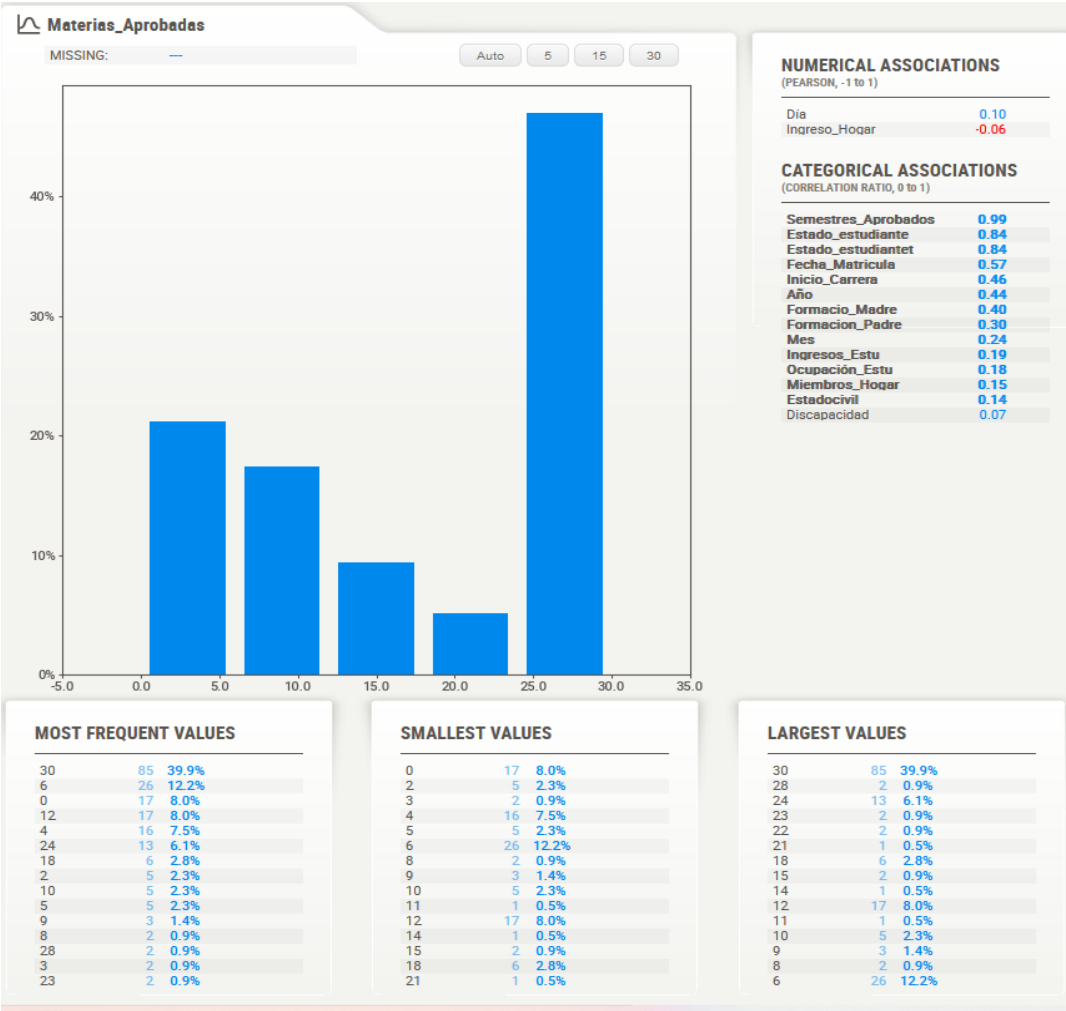


Figura 36 Datos descriptivos de la variable Materias Aprobadas
Fuente: (Cevallos, 2022)

La Figura 37 muestra datos descriptivos generales de la data set

Figura 37 Datos descriptivos de manera general
Fuente: (Cevallos, 2022)

3.3.3.3. Construcción de nuevos datos

En este punto se creará una variable fechas por separado año, mes y día, para el tiempo que ha estudiado el educando, se generará una nueva columna para tal proceso.

Figura 38 muestra el código para crear de la fecha columnas año, mes y día.

```
[229] df3['Año']=df3['Fecha_Nacimiento'].dt.year
      df3['Día']=df3['Fecha_Nacimiento'].dt.day
      df3['Mes']=df3['Fecha_Nacimiento'].dt.month
      df3['Año']=df3['Inicio_Carrera'].dt.year
      df3['Día']=df3['Inicio_Carrera'].dt.day
      df3['Mes']=df3['Inicio_Carrera'].dt.month
      df3['Año']=df3['Fecha_Matricula'].dt.year
      df3['Día']=df3['Fecha_Matricula'].dt.day
      df3['Mes']=df3['Fecha_Matricula'].dt.month
```

Figura 38 Fecha
Fuente: (Cevallos, 2022)

Figura 39 muestra las columnas creadas con año, mes y día.

```
df3.dtypes

Sexo                          int64
Estadocivil                   int64
Etnia                         int64
Discapacidad                 object
Fecha_Nacimiento     datetime64[ns]
Pais_Residencia               int64
Provincia_Residencia          int64
Cantón_Residencia             int64
Tipo_Colegio                  int64
Modalidad_Carrera             int64
Jornada_Carrera               int64
Inicio_Carrera       datetime64[ns]
Fecha_Matricula      datetime64[ns]
Repetido_Materia              int64
Ocupación_Estu                int64
Ingresos_Estu                 int64
Formacion_Padre               int64
Formacio_Madre                int64
Ingreso_Hogar               float64
Miembros_Hogar                int64
Costo_Semestre                int64
Valor_Semestre                int64
Total_Materias                int64
Materias_Semestre             int64
Semestres_Aprobados         float64
Materias_Aprobadas            int64
Estado_estudiante           float64
Estado_estudiantet           object
Año                           int64
Día                           int64
Mes                           int64
dtype: object
```

Figura 39 Tipos de datos
Fuente: (Cevallos, 2022)

Con el análisis el proceso anterior se procede a dejar solo las variables que tienen correlación alta.

Figura 40 muestra las variables de valor constante

Figura 40 Valores constantes
Fuente: (Cevallos, 2022)

Figura 41 muestra las variables de alta correlación que son las que serán tomados en cuenta para el proyecto.

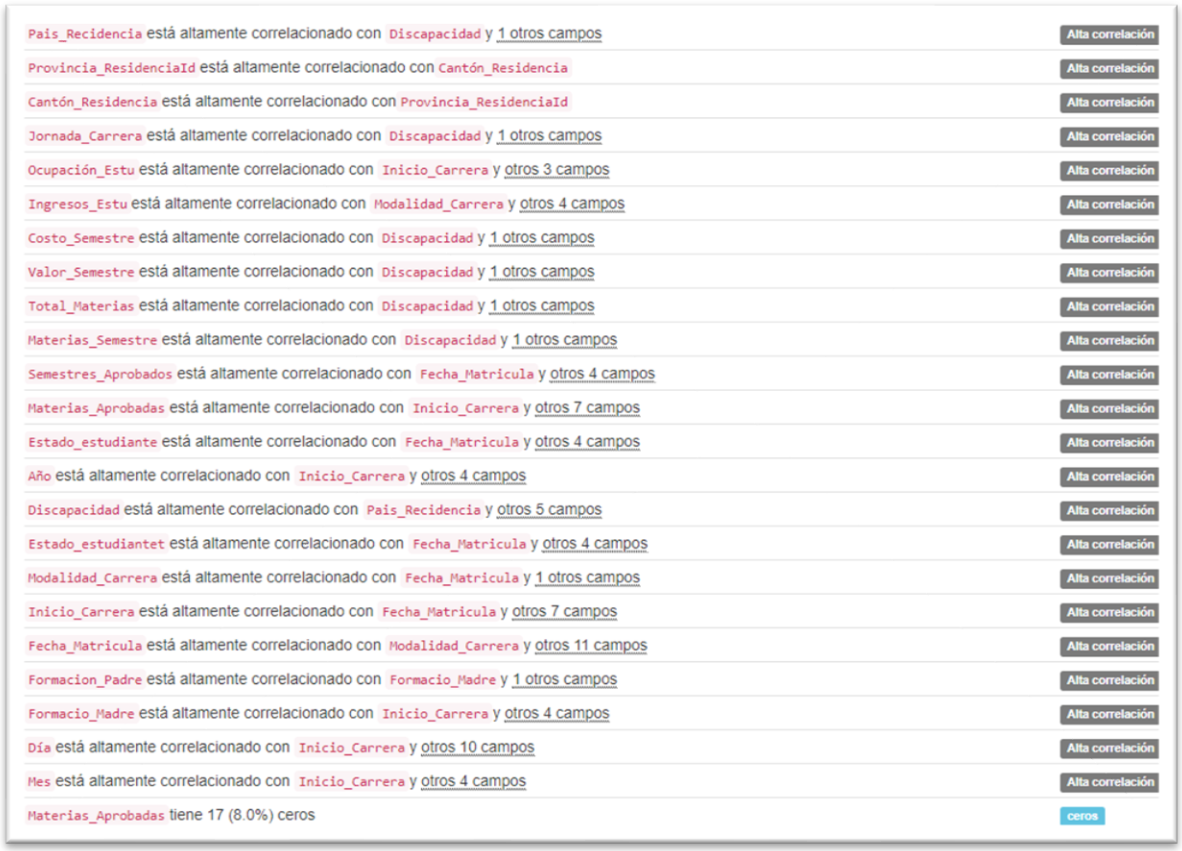


Figura 41 Valores alta correlación
Fuente: (Cevallos, 2022)

3.3.3.4. Formato de datos

La matriz de correlación nos muestra las relaciones positivas y negativas, las cuales serán tomas en cuenta para el desarrollo del proyecto

La figura 42 muestra la matriz de correlación.

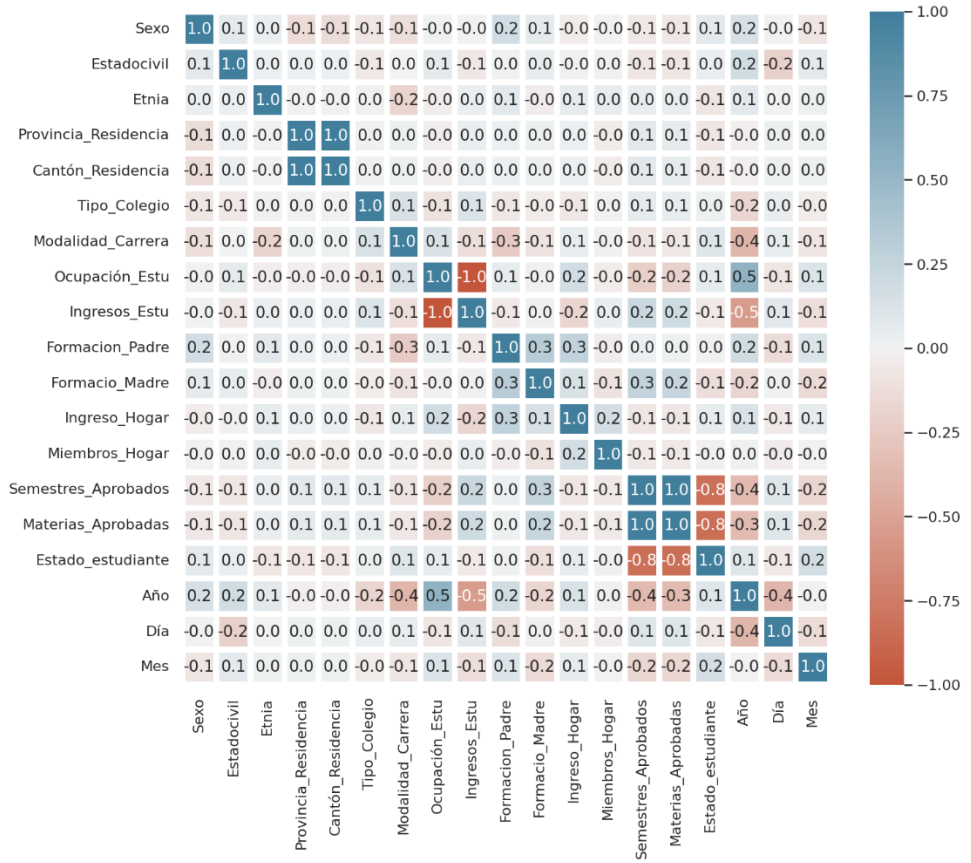


Figura 42 Matriz de correlación de Pearson
Fuente: (Cevallos, 2022)

Figura 43 muestra la matriz de correlación por colores

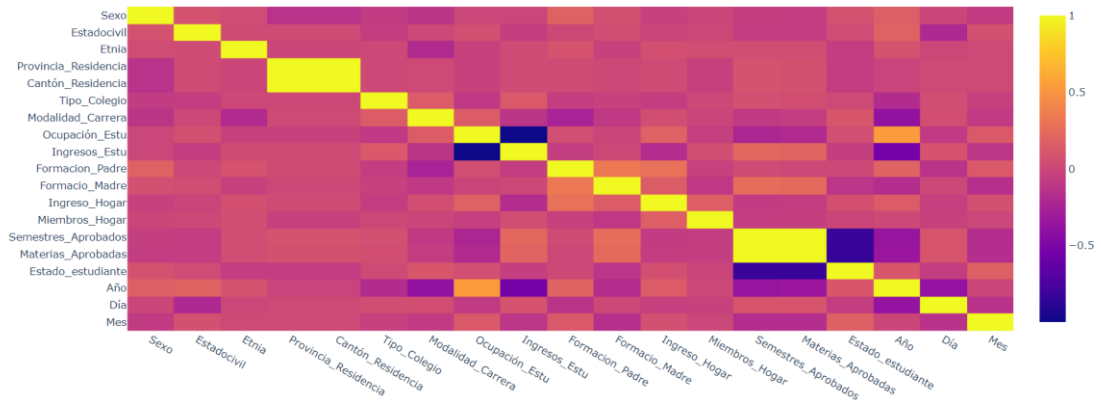


Figura 43 Matriz de correlación por colores
Fuente: (Cevallos, 2022)

Análisis de la correlación. Se observa relaciones fuertes entre variables tanto positiva o negativamente, como semestres aprobados, estado de estudiantes con estos datos se procede a elaborar varios algoritmos y así escoger el más cercano a la predicción que se busca.

3.3.4. Modelado

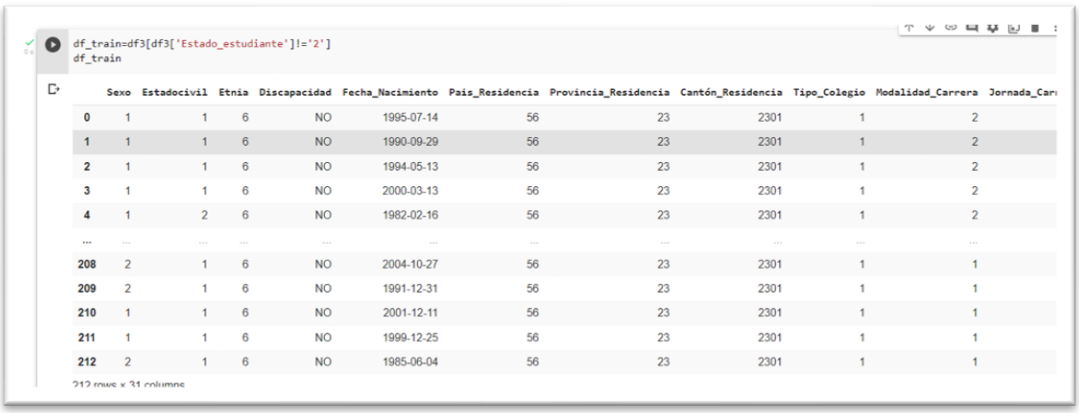
Dando cumplimiento al tercer objetivo: Aplicar métricas estadísticas para analizar las variables que ocasionan la deserción estudiantil.

3.3.4.1. Selección de técnicas de modelado

Luego de tomar las relaciones más fuertes entre variables tanto positiva o negativamente se procede a realizar el primer análisis de regresión lineal múltiple para elaborar varios algoritmos y así escoger el más cercano a la predicción que se busca.

- Se reduce los datos retirando los datos que están en estado estudiando, esté pasará a ser nuestro set de test y el otro con las deserciones y graduados serán los de entrenamiento.

Figura 44 muestra los resultados del proceso



```
df_train=df3[df3['Estado_estudiante']!='2']
df_train
```

	Sexo	Estadocivil	Etnia	Discapacidad	Fecha_Nacimiento	Pais_Residencia	Provincia_Residencia	Cantón_Residencia	Tipo_Colegio	Modalidad_Carrera	Jornada_Car
0	1	1	6	NO	1995-07-14	56	23	2301	1	2	
1	1	1	6	NO	1990-09-29	56	23	2301	1	2	
2	1	1	6	NO	1994-05-13	56	23	2301	1	2	
3	1	1	6	NO	2000-03-13	56	23	2301	1	2	
4	1	2	6	NO	1982-02-16	56	23	2301	1	2	
...	...	...	...	...	...	...	...	...	...	...	...
208	2	1	6	NO	2004-10-27	56	23	2301	1	1	
209	2	1	6	NO	1991-12-31	56	23	2301	1	1	
210	1	1	6	NO	2001-12-11	56	23	2301	1	1	
211	1	1	6	NO	1999-12-25	56	23	2301	1	1	
212	2	1	6	NO	1985-06-04	56	23	2301	1	1	

212 rows × 31 columns

Figura 44 Datos estudiantes
Fuente: (Cevallos, 2022)

3.3.4.2. Generación de un diseño de comprobación

Para regresión logística se necesita valores 1 y 0 por lo que remplazo los valores de 3 que es retirado por o y mantengo estado de estudiante en 1.

La figura 45 muestra que se reemplazó acertadamente los datos.

```
df_train['Estado_estudiante']=df_train.Estado_estudiante.replace(['1','3'],[1,3])
df_train
```

_Hogar	Costo_Semestre	Valor_Semestre	Total_Materias	Materias_Semestre	Semestres_Aprobados	Materias_Aprobadas	Estado_estudiante
5	650	5	30	6	5.0	30	1.0
0	650	5	30	6	5.0	30	1.0
8	650	5	30	6	1.0	6	3.0
5	650	5	30	6	1.0	6	3.0
5	650	5	30	6	5.0	30	1.0
...	...	...	...	...	...	...	...
7	650	5	30	6	1.0	2	3.0
1	650	5	30	6	1.0	2	3.0
3	650	5	30	6	1.0	6	2.0
1	650	5	30	6	1.0	6	2.0
2	650	5	30	6	1.0	0	3.0

Figura 45 Diseño de comprobación
Fuente: (Cevallos, 2022)

3.3.4.3. Generación y evaluación de los modelos

- Análisis de regresión simple

Se va analizar mediante la asociación de otras variables el estado estudiante.

Figura 46 muestra el modelo 1 de la regresión múltiple, Estado_estudiante vs Ingresos_Estu.

```
# MEdiante la asociacióin de otras variables ver estado estudiante
model_1 = (
    smf.ols(
        formula='Ingresos_Estu ~ Estado_estudiante',
        data=df_train
    )
    .fit()
)

model_1.summary()  #descripcion del modelo
```

Figura 46 Modelo 1 de la regresión simple
Fuente: (Cevallos, 2022)

Figura 47 muestra los resultados del modelo 1, De la variable independiente Estado_estudiante vemos que la pendiente es de -0.0537, lo que indica que por cada 1.6381 más en ingresos tiene menos probabilidad de desertar en -0.0537

OLS Regression Results

Dep. Variable:	Ingresos_Estu	**R-squared:**	0.002
Model:	OLS	**Adj. R-squared:**	-0.003
Method:	Least Squares	**F-statistic:**	0.3712
Date:	Mon, 24 Oct 2022	**Prob (F-statistic):**	0.543
Time:	15:09:28	**Log-Likelihood:**	-329.30
No. Observations:	213	**AIC:**	662.6
Df Residuals:	211	**BIC:**	669.3
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	1.6381	0.200	8.184	0.000	1.244	2.033
Estado_estudiante	-0.0537	0.088	-0.609	0.543	-0.228	0.120

Omnibus:	63.071	**Durbin-Watson:**	0.210
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	111.149
Skew:	1.706	**Prob(JB):**	7.32e-25
Kurtosis:	3.936	**Cond. No.**	6.79

Figura 47 Resultado modelo 1 de la regresión múltiple
Fuente: (Cevallos, 2022)

Figura 48 muestra el modelo 2 de la regresión múltiple

```
model_2 = (
    smf.ols(
        formula='Estado_estudiante ~ Repetido_Materia+Semestres_Aprobados',
        data=df_train
    )
    .fit()
)

model_2.summary()
```

Figura 48 Modelo 2 de la regresión múltiple
Fuente: (Cevallos, 2022)

Figura 49 muestra los resultados del modelo 2, Los errores estándar asumen que la matriz de covarianza de los errores está correctamente especificada. El valor propio más pequeño es 2.36e-28. Esto podría indicar que hay, fuertes problemas de multicolinealidad o que la matriz de diseño sea singular

OLS Regression Results

Dep. Variable:	Estado_estudiante	**R-squared:**	0.692
Model:	OLS	**Adj. R-squared:**	0.690
Method:	Least Squares	**F-statistic:**	471.3
Date:	Fri, 18 Nov 2022	**Prob (F-statistic):**	1.42e-55
Time:	01:01:46	**Log-Likelihood:**	-150.44
No. Observations:	212	**AIC:**	304.9
Df Residuals:	210	**BIC:**	311.6
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.6749	0.014	49.323	0.000	0.648	0.702
Repetido_Materia	1.3497	0.027	49.323	0.000	1.296	1.404
Semestres_Aprobados	-0.4142	0.019	-21.710	0.000	-0.452	-0.377

Omnibus:	33.769	**Durbin-Watson:**	1.609
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	52.223
Skew:	0.896	**Prob(JB):**	4.57e-12
Kurtosis:	4.643	**Cond. No.**	3.90e+15

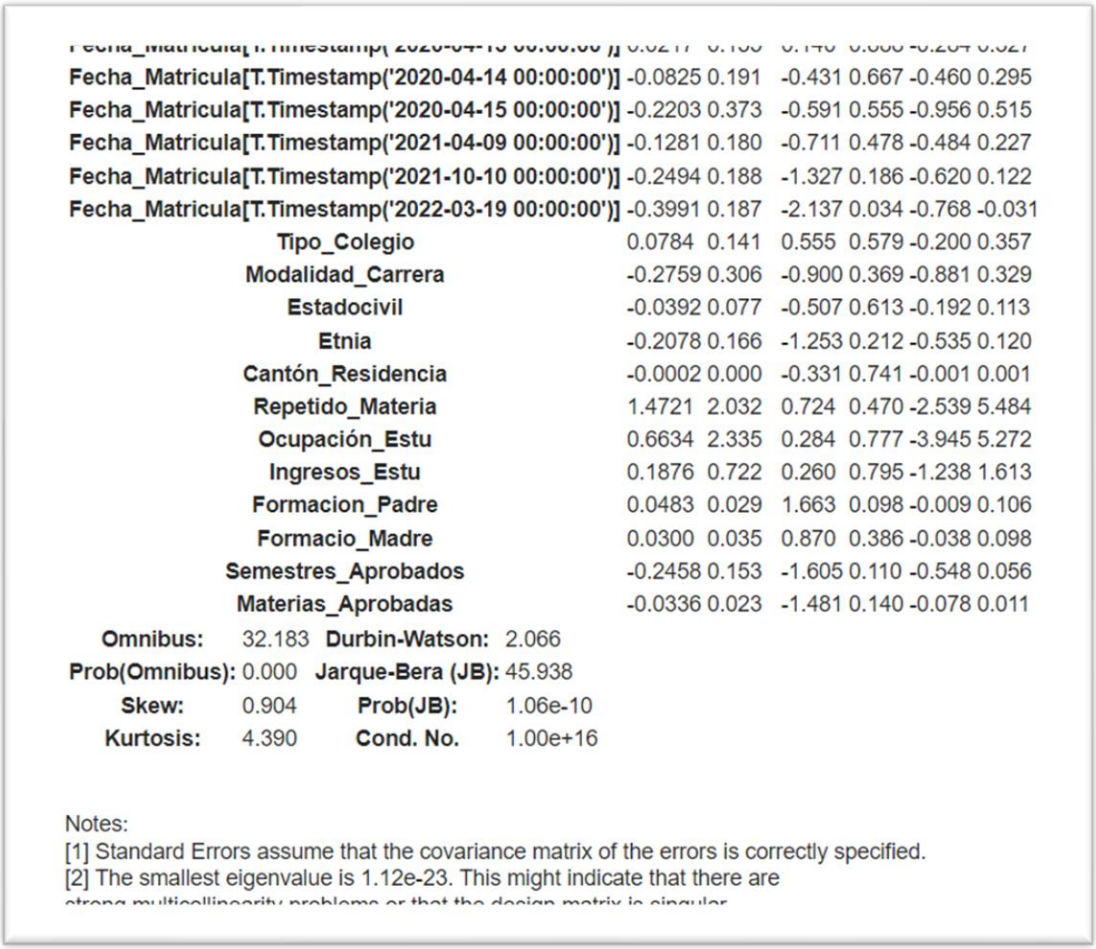
Figura 49 Resultado modelo 2 de la regresión múltiple
Fuente: (Cevallos, 2022)

Figura 50 muestra el modelo 3 de la regresión múltiple

```
model_3 = (
    smf.ols(
        formula='Estado_estudiante ~  Tipo_Colegio + Modalidad_Carrera + Estadocivil + Etnia
        + Cantón_Residencia + Inicio_Carrera+Fecha_Matricula+Repetido_Materia+Ocupación_Estu+
        Ingresos_Estu+Formacion_Padre+Formacio_Madre+Semestres_Aprobados+Materias_Aprobadas',
        data=df_train
    )
    .fit()
)

model_3.summary()
```

Figura 50 Modelo 3 de la regresión múltiple
Fuente: (Cevallos, 2022)

Figura 51 muestra los resultados del modelo 3, donde los errores estándar asumen que la matriz de covarianza de los errores está correctamente especificada. El valor propio más pequeño es 1.12e-23. Esto podría indicar que hay fuertes problemas de multicolinealidad o que la matriz de diseño sea singular



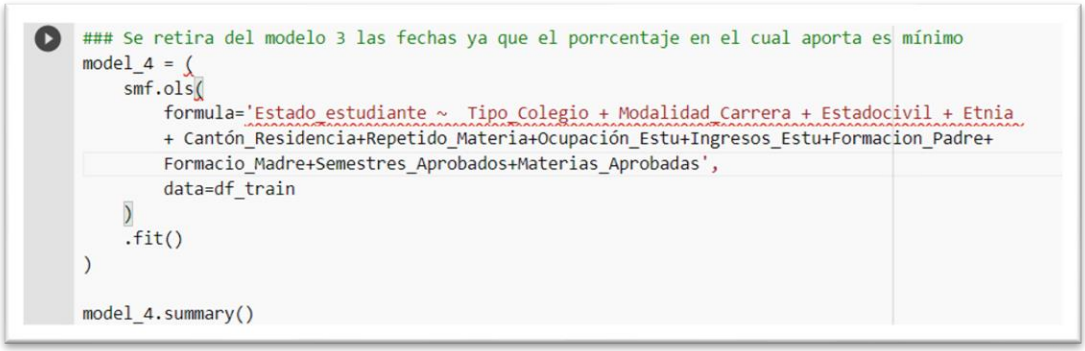
Fecha_Matricula[T.Timestamp('2020-04-14 00:00:00')]	-0.0825	0.191	-0.431	0.667	-0.460	0.295
Fecha_Matricula[T.Timestamp('2020-04-15 00:00:00')]	-0.2203	0.373	-0.591	0.555	-0.956	0.515
Fecha_Matricula[T.Timestamp('2021-04-09 00:00:00')]	-0.1281	0.180	-0.711	0.478	-0.484	0.227
Fecha_Matricula[T.Timestamp('2021-10-10 00:00:00')]	-0.2494	0.188	-1.327	0.186	-0.620	0.122
Fecha_Matricula[T.Timestamp('2022-03-19 00:00:00')]	-0.3991	0.187	-2.137	0.034	-0.768	-0.031
Tipo_Colegio	0.0784	0.141	0.555	0.579	-0.200	0.357
Modalidad_Carrera	-0.2759	0.306	-0.900	0.369	-0.881	0.329
Estadocivil	-0.0392	0.077	-0.507	0.613	-0.192	0.113
Etnia	-0.2078	0.166	-1.253	0.212	-0.535	0.120
Cantón_Residencia	-0.0002	0.000	-0.331	0.741	-0.001	0.001
Repetido_Materia	1.4721	2.032	0.724	0.470	-2.539	5.484
Ocupación_Estu	0.6634	2.335	0.284	0.777	-3.945	5.272
Ingresos_Estu	0.1876	0.722	0.260	0.795	-1.238	1.613
Formacion_Padre	0.0483	0.029	1.663	0.098	-0.009	0.106
Formacio_Madre	0.0300	0.035	0.870	0.386	-0.038	0.098
Semestres_Aprobados	-0.2458	0.153	-1.605	0.110	-0.548	0.056
Materias_Aprobadas	-0.0336	0.023	-1.481	0.140	-0.078	0.011

Omnibus:	32.183	**Durbin-Watson:**	2.066
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	45.938
Skew:	0.904	**Prob(JB):**	1.06e-10
Kurtosis:	4.390	**Cond. No.**	1.00e+16

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.12e-23. This might indicate that there are

Figura 51 Resultado modelo 3 de la regresión múltiple
Fuente: (Cevallos, 2022)

Figura 52 muestra el modelo 4 de la regresión múltiple



```
### Se retira del modelo 3 las fechas ya que el porrcentaje en el cual aporta es mínimo
model_4 = (
    smf.ols(
        formula='Estado_estudiante ~  Tipo_Colegio + Modalidad_Carrera + Estadocivil + Etnia
        + Cantón_Residencia+Repetido_Materia+Ocupación_Estu+Ingresos_Estu+Formacion_Padre+
        Formacio_Madre+Semestres_Aprobados+Materias_Aprobadas',
        data=df_train
    )
    .fit()
)

model_4.summary()
```

Figura 52 Modelo 4 de la regresión múltiple
Fuente: (Cevallos, 2022)

Figura 53 muestra los resultados del modelo 4. Los errores estándar asumen que la matriz de covarianza de los errores está correctamente especificada. El valor propio más pequeño es 4.59e-29. Esto podría indicar que hay fuertes problemas de multicolinealidad o que la matriz de diseño sea singular.

OLS Regression Results

Dep. Variable:	Estado_estudiante	**R-squared:**	0.736
Model:	OLS	**Adj. R-squared:**	0.721
Method:	Least Squares	**F-statistic:**	50.64
Date:	Fri, 18 Nov 2022	**Prob (F-statistic):**	8.54e-52
Time:	00:38:10	**Log-Likelihood:**	-134.09
No. Observations:	212	**AIC:**	292.2
Df Residuals:	200	**BIC:**	332.5
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.1489	0.824	1.394	0.165	-0.477	2.774
Tipo_Colegio	0.1673	0.120	1.394	0.165	-0.069	0.404
Modalidad_Carrera	0.1992	0.120	1.659	0.099	-0.038	0.436
Estadocivil	-0.0450	0.071	-0.634	0.527	-0.185	0.095
Etnia	-0.0916	0.161	-0.568	0.571	-0.410	0.227
Cantón_Residencia	-0.0002	0.000	-0.433	0.665	-0.001	0.001
Repetido_Materia	2.2978	1.649	1.394	0.165	-0.953	5.549
Ocupación_Estu	-0.9768	1.730	-0.565	0.573	-4.388	2.434
Ingresos_Estu	-0.2231	0.576	-0.387	0.699	-1.359	0.913
Formacion_Padre	0.0247	0.028	0.875	0.382	-0.031	0.080
Formacio_Madre	0.0605	0.028	2.152	0.033	0.005	0.116
Semestres_Aprobados	-0.0905	0.148	-0.611	0.542	-0.383	0.202
Materias_Aprobadas	-0.0534	0.022	-2.401	0.017	-0.097	-0.010

Omnibus:	21.079	**Durbin-Watson:**	1.860
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	24.316
Skew:	0.764	**Prob(JB):**	5.25e-06
Kurtosis:	3.647	**Cond. No.**	4.94e+18

Figura 53 Resultado modelo 1 de la regresión múltiple
Fuente: (Cevallos, 2022)

- Kmeans

Figura 54 muestra la data para realizar pruebas

```
[143] train_X = df_train[['Estado_estudiante','Tipo_Colegio','Modalidad_Carrera','Estadocivil','Etnia','Cantón_Residencia','Repetido_Materia','Ocupación_Estu','Ingresos_Estu','Formacion_Padre','Forma
      # taking the training data features
      train_y=df_train.Estado_estudiante
      train_X.head(2)
```

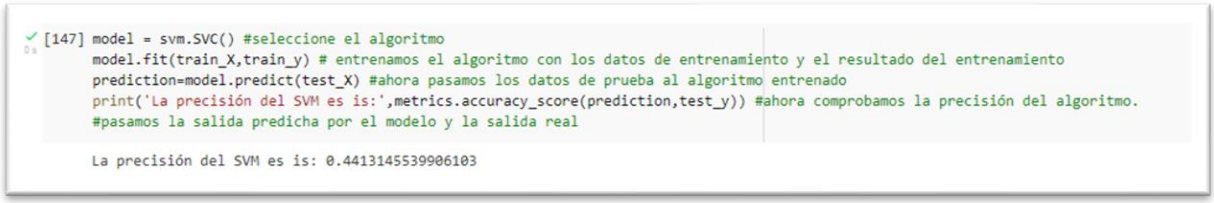
	Estado_estudiante	Tipo_Colegio	Modalidad_Carrera	Estadocivil	Etnia	Cantón_Residencia	Repetido_Materia	Ocupación_Estu	Ingresos_Estu	Formacion_Padre	Formacio_Madre	Semestres_Aprobado
0	1.0	1	2	1	6	2301	2	2	1	3	3	5
1	1.0	1	2	1	6	2301	2	2	1	3	3	5

Figura 54 Data para prueba
Fuente: (Cevallos, 2022)

Probamos en los diferentes modelos

- Modelo SVM

Figura 55 muestra resultado del modelo SVM, con una precisión del SVM es: 0.4413145539906103



```
[147] model = svm.SVC() #seleccione el algoritmo
      model.fit(train_X,train_y) # entrenamos el algoritmo con los datos de entrenamiento y el resultado del entrenamiento
      prediction=model.predict(test_X) #ahora pasamos los datos de prueba al algoritmo entrenado
      print('La precisión del SVM es is:',metrics.accuracy_score(prediction,test_y)) #ahora comprobamos la precisión del algoritmo.
      #pasamos la salida predicha por el modelo y la salida real

      La precisión del SVM es is: 0.4413145539906103
```

Figura 55 Modelo SVM
Fuente: (Cevallos, 2022)

Figura 56 muestra resultado de la precisión de la regresión logística es 0.9953051643192489, esté algoritmo es el que verificamos, se denota que existe un sobre entrenamiento por lo cual no es el más acorde para esté análisis.

```
model = LogisticRegression(max_iter=1000)
model.fit(train_X,train_y)
prediction=model.predict(test_X)
print('La precisión de la regresión logística es',metrics.accuracy_score(prediction,test_y))

La precisión de la regresión logística es 0.9953051643192489
```

Figura 56 Modelo regresión logística
Fuente: (Cevallos, 2022)

Figura 57 muestra resultado de la precisión del árbol de decisión es 1.0. La predicción de árboles de decisión muestra también sobre entrenamiento con esté set de datos, esto puede mejorar con más datos no solo los proporcionados en este caso.

```
model=DecisionTreeClassifier()
model.fit(train_X,train_y)
prediction=model.predict(test_X)
print('La precisión del árbol de decisión es',metrics.accuracy_score(prediction,test_y))

La precisión del árbol de decisión es 1.0
```

Figura 57 Modelo árbol de decisión
Fuente: (Cevallos, 2022)

Figura 58 muestra resultado de la precisión del KNN es 0.971830985915493, para esté caso de estudio el mejor es KNN por lo que se elaborará esté caso.

```
model=KNeighborsClassifier(n_neighbors=3)
model.fit(train_X,train_y)
prediction=model.predict(test_X)
print('La precisión del KNN es',metrics.accuracy_score(prediction,test_y))

La precisión del KNN es 0.971830985915493
```

Figura 58 Modelo KNN
Fuente: (Cevallos, 2022)

Para nuestro proceso usaremos regresión lineal con scikit-learn con los siguientes métodos: XGBoost, DecisionTree, RandomForest y KNeighboursRegressor el cual evaluaremos al final para ver cuál es el que mejor predice este conjunto de datos.

- ❖ Se instancian modelos de entrenamiento

Figura 59 muestra los modelos de entramiento instanciados

```
xgb = XGBRegressor()
dtc = DecisionTreeRegressor()
rfc = RandomForestRegressor()
knn = KNeighborsRegressor(n_neighbors=8)
```

Figura 59 Modelos instanciados
Fuente: (Cevallos, 2022)

❖ Se procede a entrenar los modelos

Figura 60 muestra código para entrenar modelos

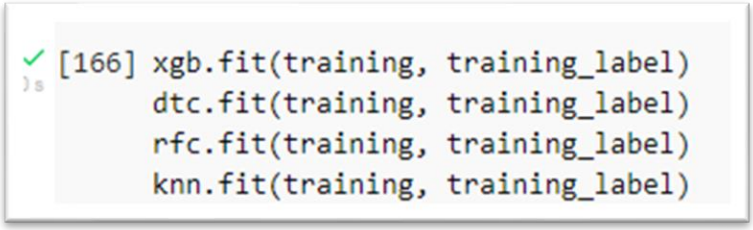


Figura 60 Entrenar los modelos
Fuente: (Cevallos, 2022)

La Figura 61 muestra código para probar nuestro modelo de predicción, donde los resultados, 'XGBoost': 0.507765152584275, 'DecisionTree': 0.5590169943749475, 'RandomForest': 0.5115824679636488, 'KNN': 0.6774838558962125, los mejores modelos son, RandomForest y DecisionTree.

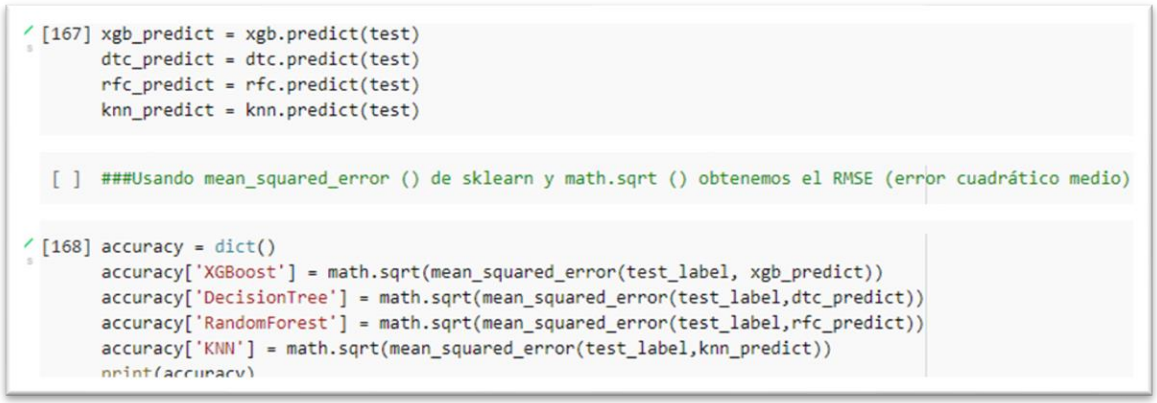


Figura 61 Probar nuestro modelo de predicción
Fuente: (Cevallos, 2022)

3.3.5. Evaluación

Dando cumplimento al cuarto objetivo: Evaluar los resultados de la predicción de la deserción de estudiantes mediante técnicas de minería de datos.

3.3.5.1. Evaluación de los resultados

❖ Resultados de la regresión múltiple

- La R cuadrado es cada vez mayor lo que indica que este último modelo captura mucho mejor el comportamiento de las variables.

- El valor de la pendiente para la variable semestres aprobados indica que a más semestres aprobados es menor la probabilidad de deserción.

❖ Resultados de los modelos SVM, Regresión logística, árbol de decisiones, KNN.

- El mejor modelo que se ajusta es el KNN, el cual se lo va a ejecutar para comprobar si predice de la mejor manera.

❖ Resultados de la regresión lineal con scikit-learn con los siguientes métodos: XGBoost, DecisionTree, RandomForest y KNeighboursRegressor.

- Se observa que RandomForest y DecisionTree tienen los números más bajos por lo que es son los algoritmos que ha realizado una mejor predicción.

3.3.6. Despliegue

3.3.6.1. Planificación de despliegue

❖ Visualización de los resultados de la regresión múltiple

Figura 62 código para mostrar los resultados de la regresión múltiple

```
models_result = pd.DataFrame(
    dict(
        actual_value = df_train.Estado_estudiante,
        prediction_model_1 = model_1.predict(),
        prediction_model_2 = model_2.predict(),
        prediction_model_3 = model_3.predict(),
        prediction_model_4 = model_4.predict(),
        Semestres_Aprobados=df_train.Semestres_Aprobados,
        Estado_estudiante=df_train.Estado_estudiante
    )
)

models_result
```

Figura 62 Código mostrar resultados regresión múltiple
Fuente: (Cevallos, 2022)

Figura 63 muestra una tabla con los valores de cada modelo con sus resultados

	actual_value	prediction_model_1	prediction_model_2	prediction_model_3	prediction_model_4	Semestres_Aprobados	Estado_estudiante
0	1.0	1.584353	1.307219	0.950471	1.256890	5.0	1.0
1	1.0	1.584353	1.307219	1.222554	1.256890	5.0	1.0
2	3.0	1.476877	2.964556	2.696598	2.913660	1.0	3.0
3	3.0	1.476877	2.964556	3.372751	3.036535	1.0	3.0
4	1.0	1.584353	1.307219	1.174391	1.335511	5.0	1.0
...	...	...	...	...	...	...	...
208	3.0	1.476877	2.964556	2.555799	2.960805	1.0	3.0
209	3.0	1.476877	2.964556	2.700577	3.035641	1.0	3.0
210	2.0	1.530615	2.964556	2.323771	2.718977	1.0	2.0
211	2.0	1.530615	2.964556	2.420290	2.768868	1.0	2.0
212	3.0	1.476877	2.964556	2.527034	3.006884	1.0	3.0

Figura 63 Resultados por modelo de la regresión múltiple
Fuente: (Cevallos, 2022)

Figura 64 grafica de los de los de regresión múltiple, donde se observa claramente el modelo 4 es el que cumple las condiciones para ser el elegido para nuestras predicciones.

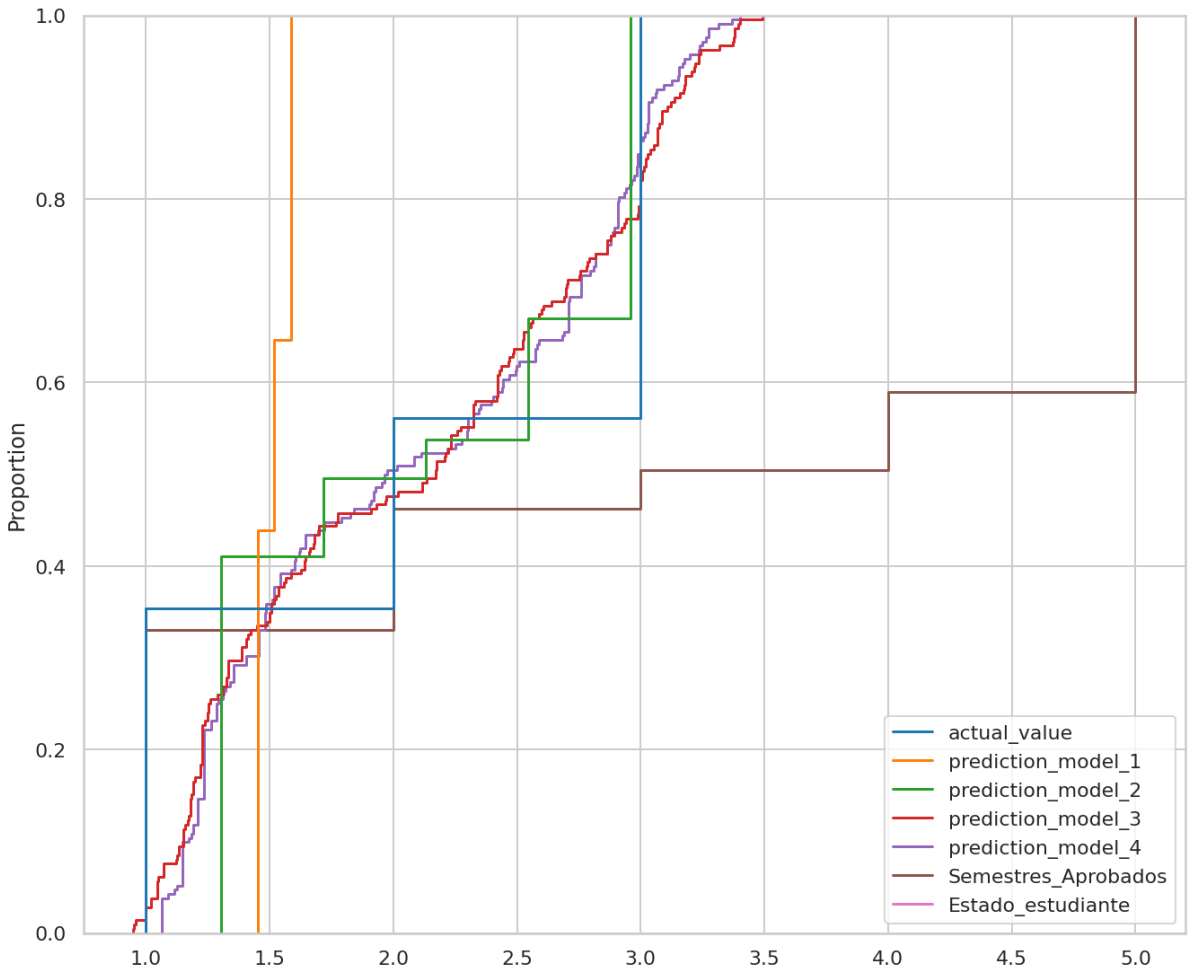


Figura 64 ECDFs de la regresión múltiple

Fuente: (Cevallos, 2022)

Figura 65 muestra solo al modelo 4 para su análisis., donde expresa que el modelo realizado en regresión múltiple no es el más adecuado por lo que realizaremos otro análisis para usar un nuevo método, para esto usaremos diferentes algoritmos que nos permitirán evaluar cual es el mejor método para este set de datos.

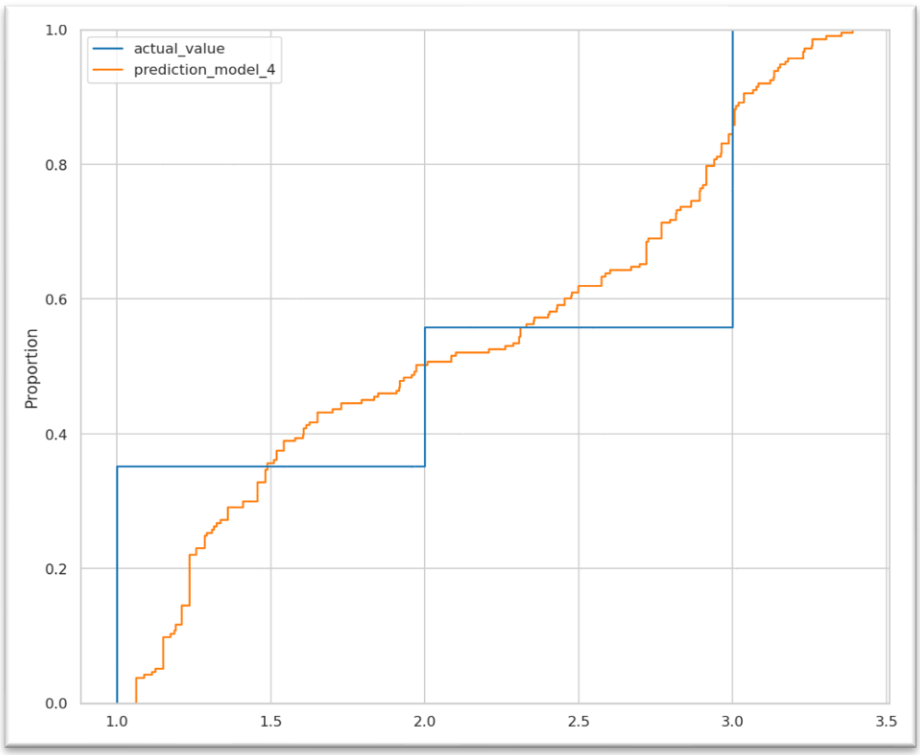


Figura 65 ECDFs del modelo 4 de la regresión múltiple
Fuente: (Cevallos, 2022)

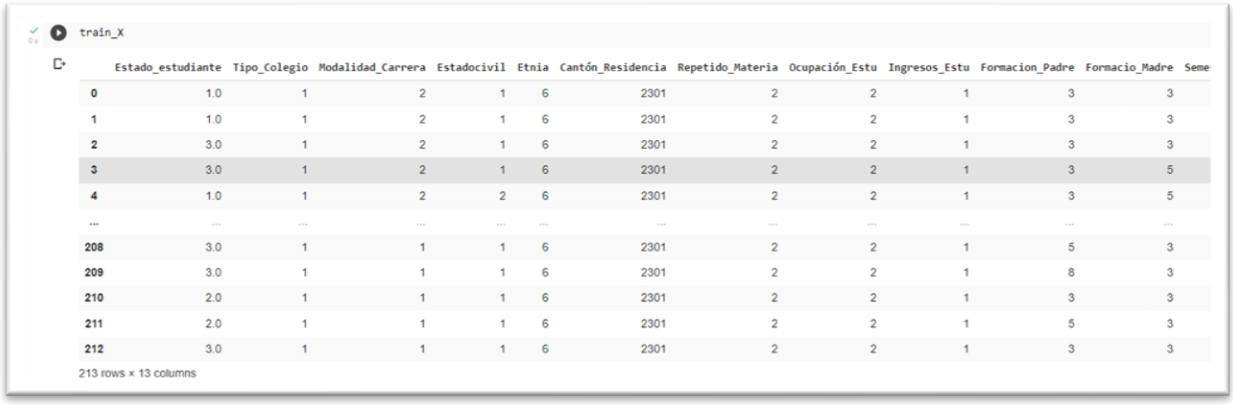
- ❖ Resultados modelo KNN

Figura 66 muestra código para ejecutar el modelo KNN

```
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
from scipy.spatial.distance import cdist

def plot_dispersion(x, figure_name, max_k = 10, n_init = 10):
    inertia = []

    for k in range(1, max_k):
        kmeans = KMeans(n_clusters = k, n_init = n_init).fit(x)
        inertia.append(kmeans.inertia_)

    plot(range(1, max_k), inertia, 'bx-')
    xlabel('k')
    ylabel(u'Dispersión')
    title(figure_name)
```

Figura 66 Código Modelo KNN
Fuente: (Cevallos, 2022)

Figura 67 muestra datos para trabajar con el modelo KNN



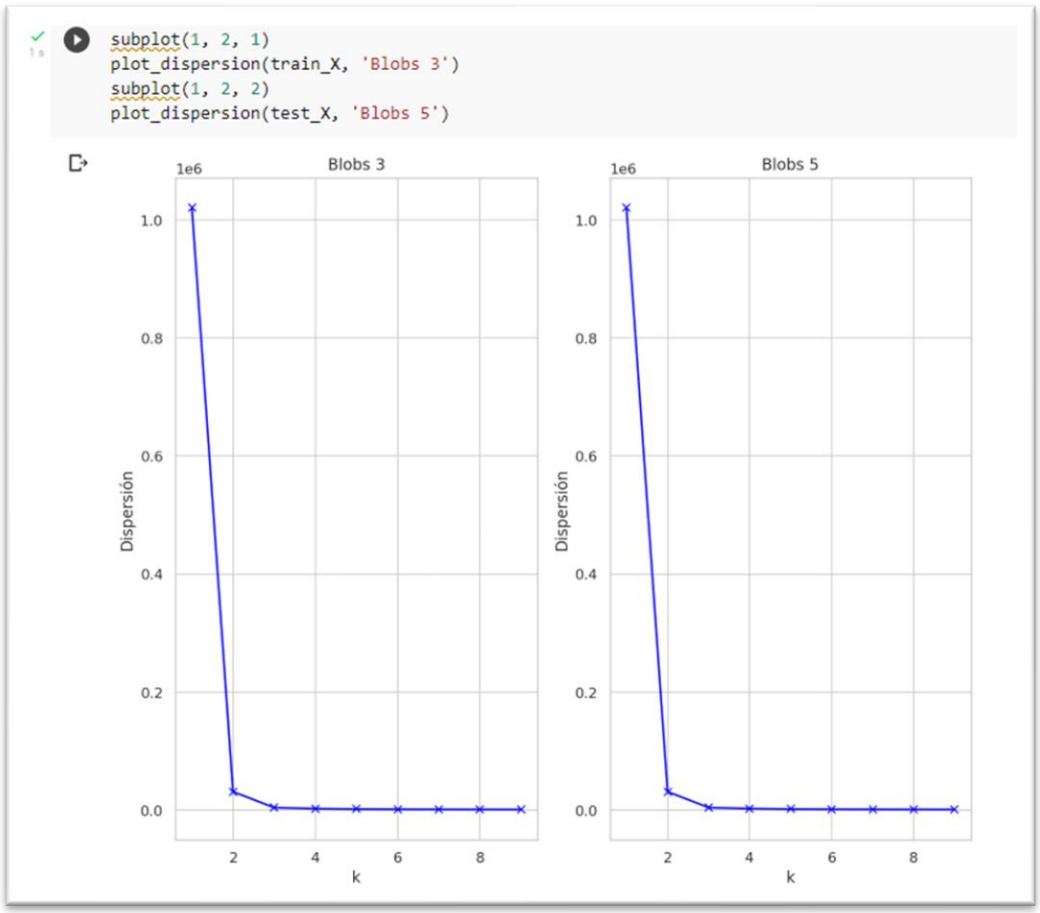
train_X

	Estado_estudiante	Tipo_Colegio	Modalidad_Carrera	Estadocivil	Etnia	Cantón_Residencia	Repetido_Materia	Ocupación_Estu	Ingresos_Estu	Formacion_Padre	Formacio_Madre	Seme
0	1.0	1	2	1	6	2301	2	2	1	3	3	
1	1.0	1	2	1	6	2301	2	2	1	3	3	
2	3.0	1	2	1	6	2301	2	2	1	3	3	
3	3.0	1	2	1	6	2301	2	2	1	3	5	
4	1.0	1	2	2	6	2301	2	2	1	3	5	
...	...	...	...	...	...	...	...	...	...	...	...	
208	3.0	1	1	1	6	2301	2	2	1	5	3	
209	3.0	1	1	1	6	2301	2	2	1	8	3	
210	2.0	1	1	1	6	2301	2	2	1	3	3	
211	2.0	1	1	1	6	2301	2	2	1	5	3	
212	3.0	1	1	1	6	2301	2	2	1	3	3	

213 rows × 13 columns

Figura 67 Datos para modelo KNN
Fuente: (Cevallos, 2022)

Figura 68 muestra los cluster con la ayuda (subplot), se puede apreciar que el número óptimo de clústeres es 8, por lo que se puede proceder al ajuste con KMeans. Una vez ajustados el modelo se puede imprimir las propiedades de los diferentes clústeres.



```
subplot(1, 2, 1)
plot_dispersion(train_X, 'Blobs 3')
subplot(1, 2, 2)
plot_dispersion(test_X, 'Blobs 5')
```


Figura 68 Cluster
Fuente: (Cevallos, 2022)

La figura 69 muestra el ajuste con KMeans

```
kmeans = KMeans(n_clusters = 8, n_init = 10).fit(train_X)
kmeans.cluster_centers_
```

```
array([[2.33333333e+00, 1.00000000e+00, 1.00000000e+00, 1.27777778e+00,
        6.00000000e+00, 2.30100000e+03, 2.00000000e+00, 1.61111111e+00,
        2.16666667e+00, 5.00000000e+00, 5.38888889e+00, 4.00000000e+00,
        2.35000000e+01],
       [3.00000000e+00, 1.00000000e+00, 1.00000000e+00, 1.00000000e+00,
        6.00000000e+00, 1.30400000e+03, 2.00000000e+00, 2.00000000e+00,
        1.00000000e+00, 4.00000000e+00, 4.00000000e+00, 1.00000000e+00,
        6.00000000e+00],
       [2.91304348e+00, 1.00000000e+00, 1.00000000e+00, 1.26086957e+00,
        6.00000000e+00, 2.30100000e+03, 2.00000000e+00, 1.82608696e+00,
        1.52173913e+00, 4.65217391e+00, 4.26086957e+00, 1.08695652e+00,
        5.65217391e-01],
       [1.16417910e+00, 1.05970149e+00, 1.10447761e+00, 1.10447761e+00,
        6.00000000e+00, 2.30100000e+03, 2.00000000e+00, 2.00000000e+00,
        1.00000000e+00, 4.50746269e+00, 4.67164179e+00, 5.00000000e+00,
        3.00000000e+01],
       [2.70370370e+00, 1.07407407e+00, 1.14814815e+00, 1.18518519e+00,
        5.88888889e+00, 2.30100000e+03, 2.00000000e+00, 1.88888889e+00,
        1.37037037e+00, 4.48148148e+00, 3.77777778e+00, 2.03703704e+00,
        1.13333333e+01],
       [2.81632653e+00, 1.04081633e+00, 1.18367347e+00, 1.16326531e+00,
        6.00000000e+00, 2.30100000e+03, 2.00000000e+00, 1.95918367e+00,
        1.12244898e+00, 4.34693878e+00, 3.91836735e+00, 1.04081633e+00,
        5.26530612e+00],
       [2.50000000e+00, 1.00000000e+00, 1.25000000e+00, 1.37500000e+00,
        6.00000000e+00, 2.30100000e+03, 2.00000000e+00, 1.87500000e+00,
        1.37500000e+00, 4.75000000e+00, 3.37500000e+00, 3.00000000e+00,
        1.72500000e+01],
       [1.20000000e+00, 1.10000000e+00, 1.00000000e+00, 1.15000000e+00,
        6.00000000e+00, 2.30100000e+03, 2.00000000e+00, 1.00000000e+00,
        4.00000000e+00, 4.40000000e+00, 4.35000000e+00, 5.00000000e+00,
        2.98000000e+01]])
```

Figura 69 Ajuste con KMeans
Fuente: (Cevallos, 2022)

La figura 70 muestra la predicción con el modelo KNN, esto según el estado del estudiante.

```
clust = kmeans.predict(train_X)

for i in range(max(clust) + 1):
    print ("Cluster", i)
    print (train_X["Estado_estudiante"][clust == i])
```

```
Cluster 0
28     3.0
34     3.0
35     2.0
43     3.0
49     3.0
58     3.0
59     3.0
116    2.0
119    2.0
120    2.0
121    2.0
122    2.0
124    2.0
125    2.0
126    2.0
171    2.0
177    2.0
178    2.0
Name: Estado_estudiante, dtype: float64
```

Figura 70 Predicción modelo KNN
Fuente: (Cevallos, 2022)

Figura 71 muestra la ejecución del algoritmo KNN.

```
from sklearn.cluster import AgglomerativeClustering

ac = AgglomerativeClustering(n_clusters = 8,
                             affinity = 'euclidean',
                             linkage = 'complete')
ac.fit_predict(train_X)
```

```
array([4, 4, 6, 6, 4, 4, 4, 4, 4, 6, 5, 6, 1, 5, 6, 6, 1, 6, 6, 0, 0, 0,
       1, 4, 0, 4, 0, 4, 2, 3, 4, 1, 4, 4, 2, 2, 4, 4, 4, 4, 4, 4, 4, 2,
       4, 3, 4, 0, 4, 2, 3, 4, 4, 4, 4, 4, 4, 4, 2, 2, 4, 1, 1, 3, 4, 4,
       6, 6, 1, 1, 4, 1, 4, 4, 4, 4, 4, 6, 4, 4, 4, 1, 4, 6, 0, 4, 4, 4,
       4, 4, 3, 4, 4, 4, 3, 4, 7, 4, 4, 4, 0, 4, 6, 4, 4, 4, 0, 4, 0, 6,
       6, 0, 4, 1, 6, 6, 2, 6, 0, 2, 2, 2, 2, 0, 2, 2, 2, 0, 0, 5, 0, 0,
       1, 1, 5, 6, 6, 5, 6, 5, 6, 1, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4, 4, 6, 0, 4, 4, 4, 4, 3, 3, 4, 0, 0, 4, 4, 6, 2, 6, 4, 6, 4,
       4, 2, 2, 6, 4, 6, 1, 0, 6, 1, 0, 1, 6, 1, 1, 1, 1, 1, 0, 0, 1, 0,
       6, 6, 6, 6, 6, 6, 6, 0, 6, 6, 0, 0, 6, 6, 0])
```

Figura 71 Algoritmo KNN.
Fuente: (Cevallos, 2022)

Luego de analizar kmeans, también se observa que no es un algoritmo que tendrá una predicción acertada por lo que se evaluarán otros algoritmos.

La figura 72 muestra el rendimiento de los diversos algoritmos, para poder visualizar lo expuesto como mejor predicción.

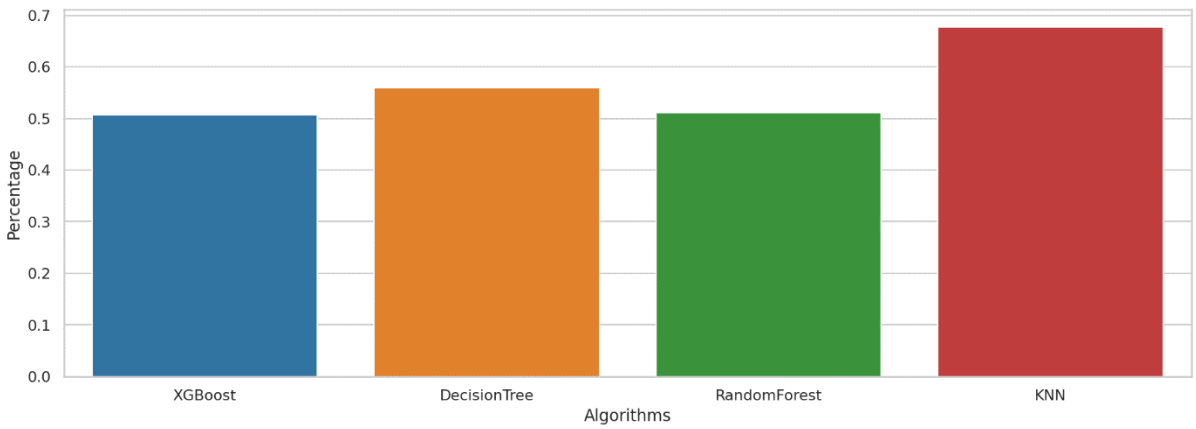
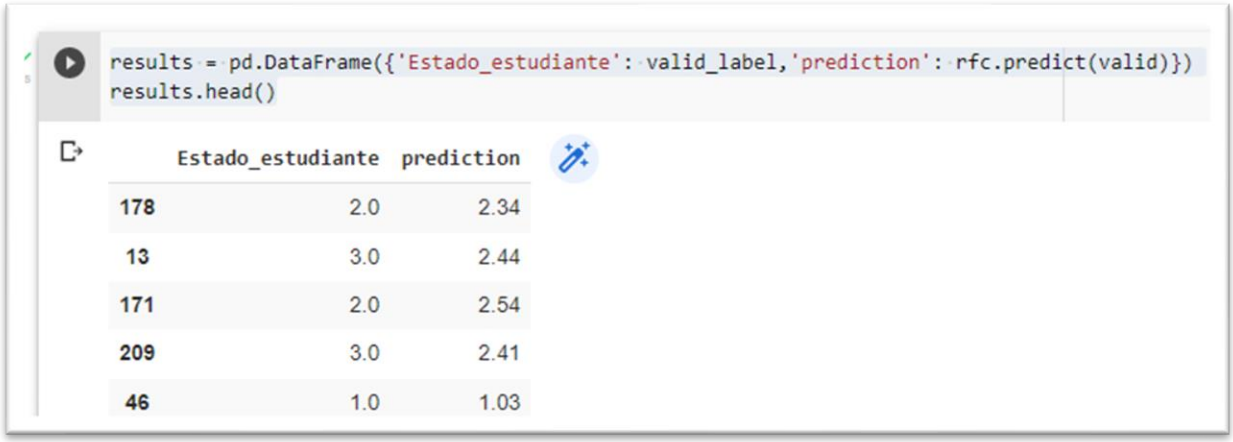


Figura 72 Algoritmos de predicción
Fuente: (Cevallos, 2022)

De todos los modelos realizado y analizados se concluye que el RandomForest es el que mejor se ajusta al proyecto y que mejores resultados arrojo, como muestras la figura, se procede a imprimir los resultados de la predicción de deserción estudiantil mediante RandomForest.

Figura 73 muestra la predicción de la deserción estudiantil



```
results = pd.DataFrame({'Estado_estudiante': valid_label,'prediction': rfc.predict(valid)})
results.head()
```

	Estado_estudiante	prediction
178	2.0	2.34
13	3.0	2.44
171	2.0	2.54
209	3.0	2.41
46	1.0	1.03

Figura 73 Predicción
Fuente: (Cevallos, 2022)

4. CONCLUSIONES Y RECOMENDACIONES

4.1. CONCLUSIONES

En el estudio se evidenció que los estudiantes del Instituto Superior Tecnológico Los Andes (ISTLA), tienen diferentes características socioeconómicas que influyen de una u otra manera en la deserción estudiantil, siendo la situación económica del núcleo familiar una de ellas teniendo ingresos desde los 600 dólares hasta 2000 dólares, además, en la mayoría de casos cuentan con más de 3 miembros en el hogar, de igual forma, otro de los factores identificados es que cuentan con un núcleo familiar (padres) sin estudios de tercer nivel, inclusive en la mayoría de los casos sin estudios de primaria finiquitados.

Se utilizó countplot para analizar datos de las variables más relevantes y con la ayuda de lamba se evidencia gráficamente la proporción de valores nulos por cada variable, además, para obtener datos estadísticos se utilizó la herramienta sweetvizcon y con la información proporcionada se eliminaron variables no relevantes o que no tenían ningún valor agregado.

Se utilizo la metodología CRISP DM, aplicando técnicas de minería de datos, se revisó proporción del estado del estudiante para ver la cantidad de graduados, estudiando y retirados. Con el coeficiente de correlación de Pearson, se analizó las variables en estudio, observando una relación fuerte entre variables ingreso en el hogar, nivel de formación del padre y de la madre y estado del estudiante, evidenciando así algunas de las variables que ocasionan la deserción estudiantil.

Se aplicaron métodos de aprendizaje automático. generando modelos capaces de predecir la deserción estudiantil. Obteniendo escenarios de regresión logística, redes neuronales, KNN y random forest, dando como mejor modelo el random forest, generando la mejor predicción, con un margen de error de 0.318382.

4.2. RECOMENDACIONES

La selección de la técnica de minería de datos debe ir afín con los objetivos del proyecto y con el tipo de datos que se manejarán para la elaboración del modelo. En el presente estudio se evidenció los problemas socioeconómicos de los estudiantes dentro de la institución, por consiguiente, la misma debe tomar medidas de seguimiento de los educandos con la finalidad de proponerles acciones de mejora a su situación puede ser manera incentivos económicos, ayuda psicológica entro otras.

Para la fase de preparación de datos es aconsejable utilizar un software que facilite la limpieza con el propósito de eliminar datos incompletos, erróneos o duplicados. Estas herramientas las debe aplicar con la finalidad de identificar las variables que están influyendo en la deserción estudiantil, la aplicación debe ser gradual y planificada.

Es importante que las autoridades del Instituto Superior Tecnológico los Andes - ISTLA tengan en consideración los resultados obtenidos en esta investigación con la finalidad de tomar decisiones y plantear estrategias que mejoren la permanencia de los alumnos, revisando constantemente la variable de estado del estudiante, con esto prevenir y reducir el índice de deserción estudiantil actual.

Se debe seguir generando y probando modelos que ayuden a predecir la deserción estudiantil. Con una mayor información e integración de datos puedan servir para proyectos futuros de Minería de Datos y Inteligencia de negocios, facilitando a los administrativos una mejor toma de decisiones.

4.3. LINEAS DE TRABAJO FUTURO

Como planes futuros se propone la creación nuevos modelos, compilando información sobre formas de aprendizaje, variables institucionales y relaciones interpersonales de los estudiantes para conseguir un modelo que permita descubrir nuevos patrones que influyen en la deserción estudiantil. Es necesario darle la importancia del caso a este proyecto de investigación, orientando los procesos de investigación a otras problemáticas del instituto. La institución debe facilitar más información y con esto poder realizar una predicción más adecuando con una data set con toda la información necesaria.

5. REFERENCIAS BIBLIOGRÁFICAS

Álvarez, C. (22 de 04 de 2020). España es el país con mayor tasa de abandono escolar de la UE. Obtenido de https://www.elperiodico.com/es/sociedad/20200422/espana-pais-mayor-tasa-abandono-escolar-ue-educacion-7936724

Bedregal, A. N., Aruquipa, V. D., & Cornejo, A. V. (marzo de 2020). Técnicas de Data Mining para extraer perfiles comportamiento académico y predecir la deserción universitaria. Obtenido de https://cutt.ly/jNuoixp

Calderón, Á., Dini, M., & Stumpo, G. (2018). Los desafíos del Ecuador para el cambio estructural con inclusión social. https://www.cepal.org/sites/default/files/publication/files/40863/S1601309_es.pdf: CEPAL.

Castillo, M. d. (2012). Causas, consecuencias y prevención de la deserción escolar. https://www.redalyc.org/pdf/654/65456040007.pdf: EUA.

CEACES, (2021), Instructivo carga masiva sistema de información integral de la educación superior: https://www.caces.gob.ec/

Corzo, C. (2020). Deserción Escolar. https://cutt.ly/UNuorPG

Contreras, B. L., Rodríguez, M. J., & Fuentes, L. H. (11 de enero de 2021). Analìtica Academica: Nuevas herramientas aplicadas a la educaciòn. Obtenido de https://dialnet.unirioja.es/servlet/articulo?codigo=792562

CEUPE. (s.f. de s.f. de 2022). *CEUPE Magazine*. Obtenido de Proceso del Data Mining: https://cutt.ly/QNuy0z4

Dominguez, C. J. (2018). Introducción al Modelado de datos. Venezuela: IEASS.

Espíndola, E., & León, A. (2002). La deserción escolar en América Latina:. Obtenido de un tema prioritario para la agenda regional: https://rieoei.org/historico/documentos/rie30a02.htm

Ethem, A. (septiembre de 2014). Introducción al aprendizaje automático.

Obtenido de https://www.cmpe.boun.edu.tr/~ethem/i2ml3e/

Hernández, Á., & Aranda, J. (2017). El Problema de la Deserción Escolar en la Producción Científica Educativa. https://www.redalyc.org/pdf/654/65456040007.pdf: Revista Internacional de Ciencias.

Hernández, R., Fernández, C., & Baptista, L. (2014). *Metodología de la Investigación* (Punta Santa Fe C.P. 01376 Sexta Edición ed.). México: McGRAW-HILL / INTERAMERICANA.

IBM. (2021). *Conceptos básicos de ayuda de CRISP-DM*. Obtenido de https://cutt.ly/hNuurCb

BM. (17 de agosto de 2021). *IBM SPSS Modeler*. Obtenido de IBM: https://cutt.ly/GNuy5pa

Joyanes, A. L. (2019). *Inteligencia de negocios y analítica de datos: Una visión global de Business Intelligence & Analytics.* Bogota: Alpha Editorial.

Jiménez, T. J., & Timarán, P. S. (Diciembre de 2015). Caracterización de la deserción estudiantil en educación superior con mineria de datos. Obtenido de http://200.10.150.204/index.php/tecnologica/article/view/453/318

Lander, E. (2000). La colonialidad del saber: eurocentrismo y ciencias sociales. Perspectivas latinoamericanas. CLACSO.

Martínez, J., & Ortega, A. (2020). La problemática actual de la deserción escolar, un análisis desde lo local. https://cutt.ly/BNui5j2

Miranda, M. A., & Guzmán, J. (junio de 2017). Análisis de la Deserción de Estudiantes Universitarios usando Técnicas de Minería de Datos. Obtenido de https://cutt.ly/XNuodye

Moreno, M., Ortiz, Y., & González, M. (2016). Capacitación de docentes en procesos neurocognitivos para atender la deserción escolar en procesos neurocognitivos para atender la deserción escolar asociada a aprovechamiento académico. Revista Puertorriqueña de Psicología.

Moreau, D. (diciembre de 2019). La ética como problema filosófico de la educación. Obtenido de https://publons.com/publon/33268502/

Muñoz, B. C. (2013). Deserción escolar, un concepto que no concluye: casos de no conclusión satisfactoria del ciclo escolar en la Institución educativa de Santa Librada. Obtenido de https://cutt.ly/8NuiLBy

Ortega, P., Macías, M., & Hernández, M. (2016). Causas de la deserción escolar e la telesecundaria de la zona 55. Revista Huella de la Palabra.

Páramo, & Correa. (2012). “Deserción estudiantil universitaria. Conceptualización. Revista Universidad EAFIT.

Ruiz, R., García, J., & Pérez, M. (2016). Causas y Consecuencias de la Deserción escolar en el bachillerato: caso Universidad Autónoma de Sinaloa. ISSN: 1665-0441: Ra Ximhai.

Santos, R. M., Mella, N. I., & García, Á. J. (septiembre de 2021). Educación moral y ética de la acción en el aprendizaje-servicio universitario: La sombra de John Dewey. Obtenido de https://perfileseducativos.unam.mx/iisue_pe/index.php/perfiles/article/view/59818

Torres, Z. C., Arce, R. C., & Lam, M. J. (junio de 2016). Estudio de variables que influyen en la deserción de estudiantes universitarios de primer año, mediante minería de datos. Obtenido de https://cutt.ly/INuiNC5

UNESCO. (2016). *Desglosar el Objetivo de Desarrollo Sostenible 4 Educación 2030.* Obtenido de https://cutt.ly/7NuuwBc

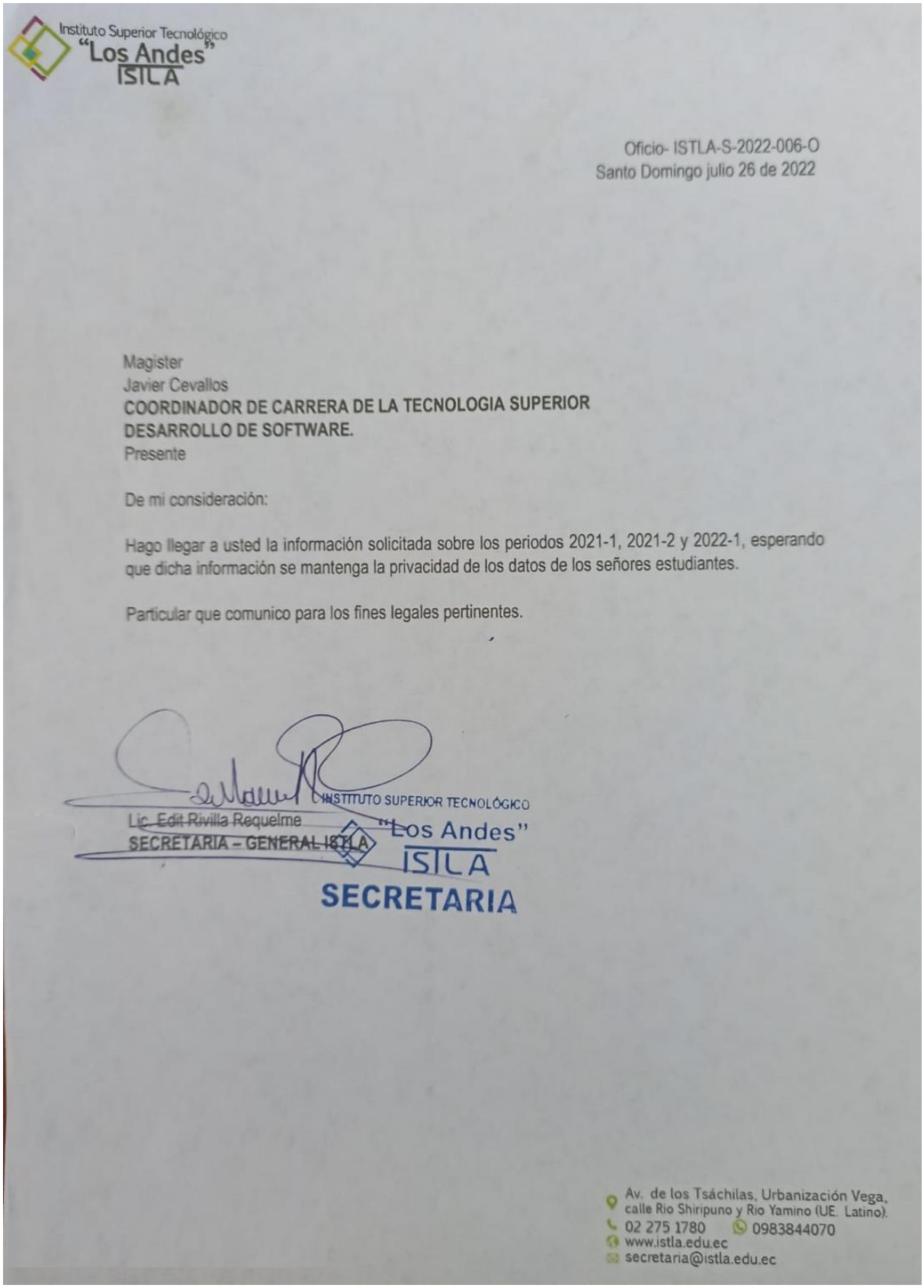
UNESCO. (22 de septiembre de 2020). La formación técnica como respuesta bisagra para reducir la potencial deserción causada por el Coronavirus. Obtenido de https://es.unesco.org/news/formacion-tecnica-como-respuesta-bisagra-reducir-potencial-desercion-causada-coronavirus

Vieira, B. L., Ortíz, V. L., & Ramírez, C. S. (2009). Introducción a la Minería de Datos. Rio de Janeiro: Editora E-papers.

UNIVERSO. (2021). La pandemia en Ecuador provoca más de 90 mil deserciones escolares. https://www.eluniverso.com/noticias/ecuador/en-ecuador-90-mil-estudiantes-dejaron-de-asistir-a-clase-durante-la-pandemia-nota/.

6. ANEXOS

Anexo 1 Documento de entrega de datos para el proyecto



Instituto Superior Tecnológico "Los Andes" ISTLA

Oficio- ISTLA-S-2022-006-O
Santo Domingo julio 26 de 2022

Magister
Javier Cevallos
COORDINADOR DE CARRERA DE LA TECNOLOGIA SUPERIOR DESARROLLO DE SOFTWARE.
Presente

De mi consideración:

Hago llegar a usted la información solicitada sobre los periodos 2021-1, 2021-2 y 2022-1, esperando que dicha información se mantenga la privacidad de los datos de los señores estudiantes.

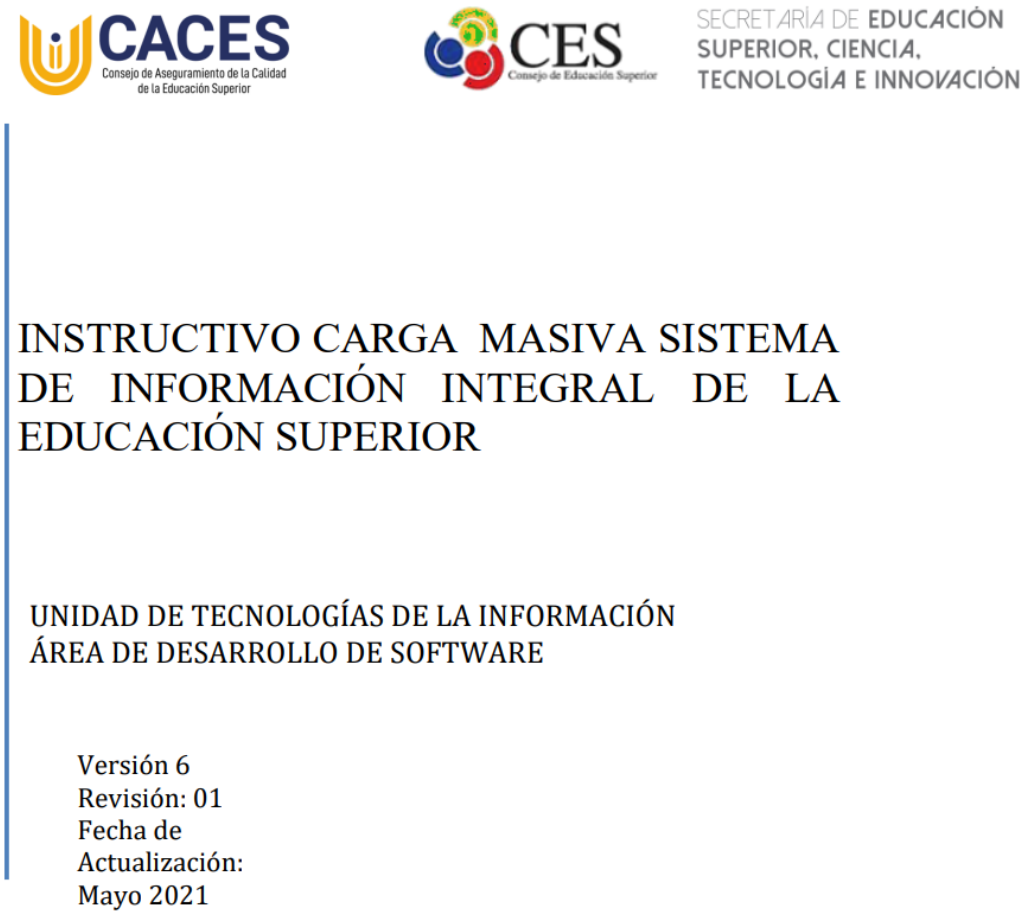
Particular que comunico para los fines legales pertinentes.

Lic. Edit Rivilla Requelme
SECRETARIA – GENERAL ISTLA

INSTITUTO SUPERIOR TECNOLÓGICO "Los Andes" ISTLA SECRETARIA

Av. de los Tsáchilas, Urbanización Vega, calle Rio Shiripuno y Rio Yamino (UE. Latino).
02 275 1780 0983844070
www.istla.edu.ec
secretaria@istla.edu.ec

Anexo 2 Documento del CACES descripción de las variables



INSTRUCTIVO CARGA MASIVA SISTEMA DE INFORMACIÓN INTEGRAL DE LA EDUCACIÓN SUPERIOR

UNIDAD DE TECNOLOGÍAS DE LA INFORMACIÓN
ÁREA DE DESARROLLO DE SOFTWARE

Versión 6
Revisión: 01
Fecha de Actualización:
Mayo 2021

Fuente: (CACES, 2021)

Enlace para visualizar al archivo completo: https://cutt.ly/wNutBxL

Anexo 3 Cronograma de actividades

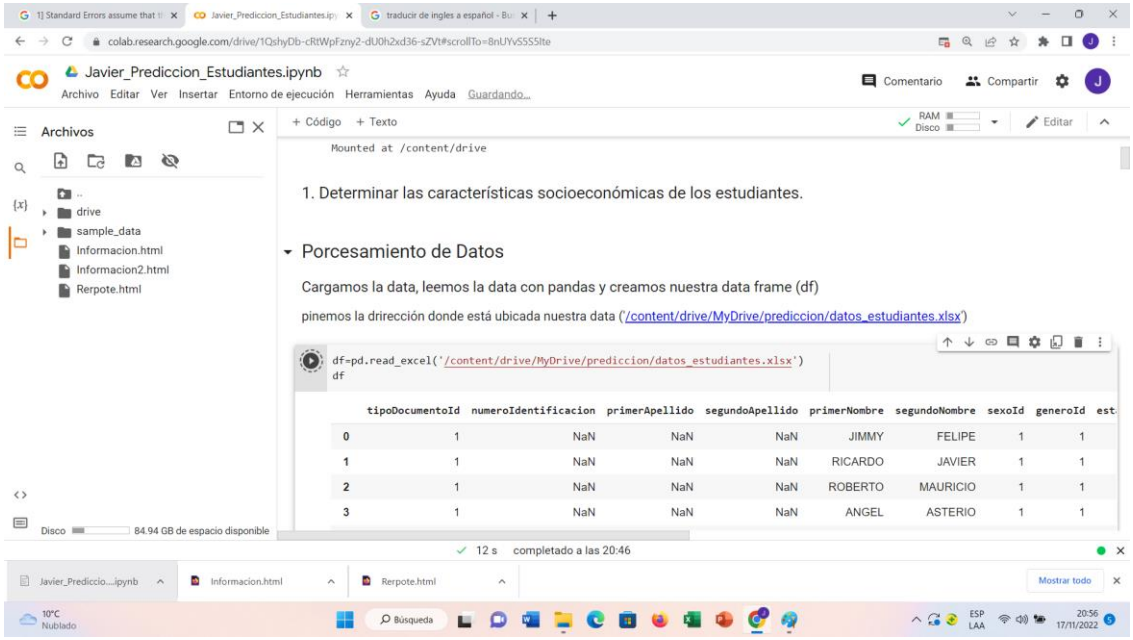
N.-	CRONOGRAMA	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC
1	Desarrollo del Plan									
2	Datos Generales									
3	Resumen Ejecutivo									
4	Descripción detallada de la propuesta									
5	Cronograma									
6	Bibliografía									
7	Anexos									
8	Elaboración del informe final de tesis									
9	Disertación del grado									

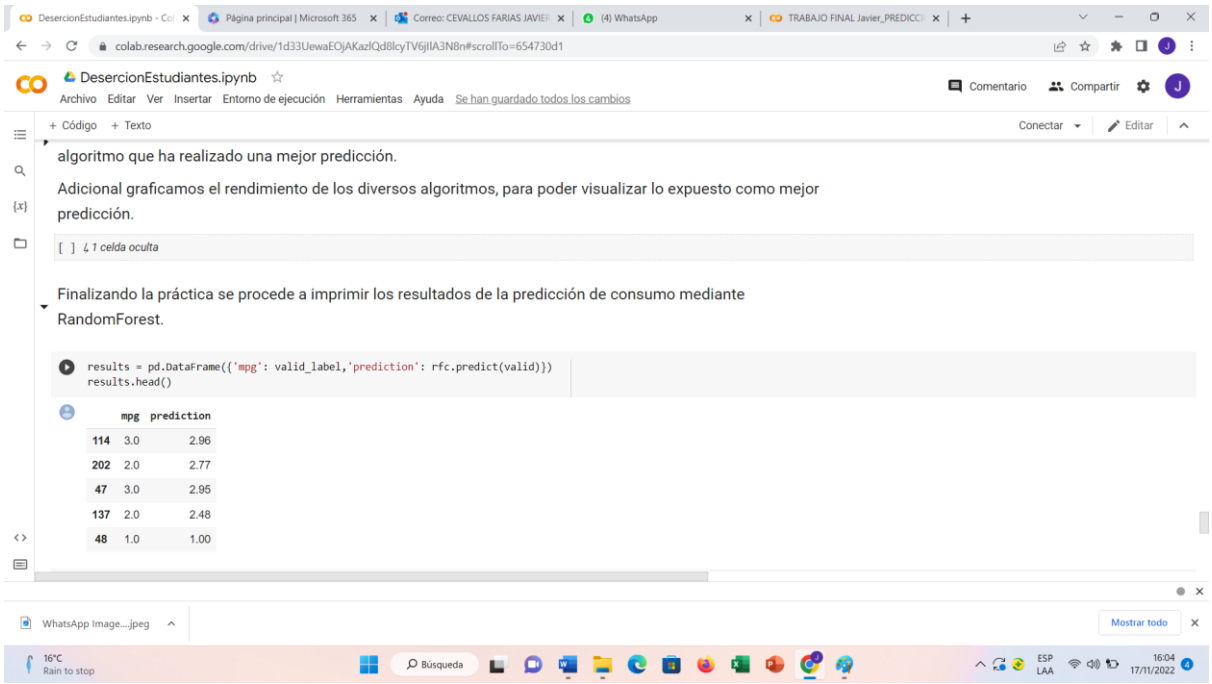
Fuente: (Cevallos, 2022)

Anexo 4 Cuaderno del proyecto (Google colab)

Enlace del cuaderno Google colab:

https://drive.google.com/drive/folders/1Gfl7gxznA_RoUPDgHTKuRtR8eeYjPcAw?usp=sharing





Fuente: (Cevallos, 2022)

Anexo 5 Fotos de la institución





Fuente: (Istla, 2022)