PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

MAESTRÍA EN SISTEMAS DE INFORMACIÓN MENCIÓN DATA SCIENCE

Modelo predictivo de los factores de deserción escolar de las instituciones educativas en bachillerato en el Ecuador de los años 2012 – 2021.

Trabajo de titulación previo a la obtención del título de Magíster en Sistemas de Información mención Data Science

Línea de investigación: Adquisición, preprocesamiento, gestión y gobernanza de datos

Autora:
María José Melo Aguilar

Tutor del trabajo de titulación
Dr. Rafael Melgarejo

Quito – Ecuador
Agosto, 2024

ÍNDICE DE CONTENIDO

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE TABLAS

Dedicatoria

Dedico mi investigación a mi familia por su apoyo incondicional durante mi vida universitaria, y a mi pareja por creer en mí y apoyarme aún en los momentos más difíciles, han sido mi fuente de inspiración para continuar superándome día a día.

1. CAPÍTULO I

1.1. Resumen ejecutivo

La presente investigación se enfoca en determinar los factores que causan la deserción escolar en el nivel educativo medio (bachillerato) en el Ecuador, la cual se ve exacerbada por factores socioeconómicos, demográficos, políticas, ente otros. La presente investigación busca determinar estos factores de deserción escolar con datos a nivel de instituciones educativas, mediante la aplicación de técnicas de minería de datos.

El objetivo de la investigación es elaborar un modelo predictivo de deserción escolar identificando los factores que llevan a que los estudiantes de bachillerato abandonen sus estudios. Para ello, el estudio se basó en la metodología *Knowledge Discovery in Databases (KDD*, por sus siglas en inglés) siguiendo las etapas de selección, preprocesamiento, transformación, minería de datos y evaluación. Se realizó un modelado de datos mediante algoritmos de aprendizaje supervisado de clasificación, la cual es la recomendada de acuerdo con las investigaciones académicas sobre modelos predictivos de deserción escolar.

Se identificó los factores que influyen en la deserción de estudiantes del nivel educativo medio (bachillerato) que comprende 1ro, 2do y 3er año de bachillerato de los períodos escolares 2012-2021 en el Ecuador, para posteriormente validar y evaluar los resultados obtenidos. Las variables utilizadas provienen de los registros administrativos del Ministerio de Educación. Las variables son; tipo de educación, nivel de educación, el área geográfica de la escuela, el régimen escolar, sostenimiento de la institución, número de estudiantes matriculados, número de docentes por escuela, entre otras. La modelización consistió en elaborar algoritmos de regresión logística, árboles de decisión, y otros modelos de clasificación para predecir los factores de deserción escolar, evaluar el desempeño de los modelos mediante métricas de evaluación y determinar el mejor modelo predictivo.

Esta investigación contribuirá a una comprensión de los factores que determinan la deserción escolar en una determinada institución educativa en el Ecuador y proporcionará un modelo predictivo que en el futuro permitirá a los tomadores de decisiones del sector educativo mejorar la oferta educativa y tomar medidas de detección temprana de la deserción.

Palabras clave: ciencia de datos, modelo predictivo, modelos de clasificación, data mining, data science, deserción escolar, Ecuador, bachillerato, educación, Covid-19.

1.2. Executive summary

This research focuses on determining the factors that cause school dropout at the middle school level (high school) in Ecuador, which is exacerbated by socioeconomic, demographic and political factors, among others. The present study seeks to determine these school dropout factors with institutional level data, through the application of data mining techniques.

The objective of the research is to develop a predictive model of school dropout by identifying the factors that lead high school students to abandon their studies. For this purpose, the study was based on the Knowledge Discovery in Databases (KDD) methodology following the stages of selection, preprocessing, transformation, data mining and evaluation. Data modeling was performed using supervised learning algorithms for classification, which is recommended according to academic research on dropout predictive models.

The factors that influence the dropout of students at the middle level of education (high school) comprising 1st, 2nd and 3rd year of high school for the 2012-2021 school periods in Ecuador were identified, in order to subsequently validate and evaluate the results obtained. The variables used come from the administrative records of the Ministry of Education. The variables are: type of education, level of education, geographical area of the school, school regime, institution maintenance, number of students enrolled, number of teachers per school, among others. The modeling consisted of developing logistic regression algorithms, decision trees, and other classification models to predict dropout factors, evaluating the performance of the models using evaluation metrics, and determining the best predictive model.

This research will contribute to an understanding of the factors that determine school dropout in a given educational institution in Ecuador and will provide a predictive model that in the future will allow decision makers in the educational sector to improve the educational offer and take measures for early detection of dropout.

Key words: data science, predictive model, classification models, data mining, data science, school dropout, Ecuador, high school, education, Covid-19.

1.3. Contextualización del tema

La deserción escolar en un grave problema socioeducativo por el que la región de América Latina y el Caribe se ve afectada desde hace ya varios años. Este fenómeno que aqueja los sistemas educativos latinoamericanos se vio exacerbado por la pandemia de Covid-19, declarada como tal el 26 de febrero de 2020. Como resultado de este fenómeno sanitario, muchos niños, niñas y adolescentes dejaron de asistir a la escuela, interrumpiendo así sus estudios y truncando sus posibilidades de un mejor futuro. Existen varios factores que explican este problema, siendo uno de los más evidentes el factor económico, pues quienes pertenecen a los estratos más pobres se encuentran en mayor vulnerabilidad, por ende, sus probabilidades de desertar son mayores. En efecto, la evidencia a más de 20 años de que se iniciaran los procesos de reforma educativa en América Latina muestra como el abandono y la deserción escolar permanece afectando principalmente a los estudiantes más pobres y vulnerables de las distintas sociedades (Román, 2013).

A inicios de la pandemia por Covid-19, los países y territorios de América Latina y el Caribe se acogieron a las recomendaciones de la Organización Mundial de la Salud (OMS) de cerrar todas las escuelas para evitar posibles contagios, ya que en la actualidad las aulas de clase son consideradas un foco de contagio para enfermedades comunes. Por ello, los países tuvieron que buscar alternativas para continuar brindando sus servicios educativos y también medidas para recuperar los aprendizajes que se perdieron durante el cierre escolar.

A inicios de febrero de 2021, este cierre de los centros educativos afectó a 124 millones de niños, niñas y adolescentes en toda la región (Fondo de las Naciones Unidas para la Infancia, UNICEF, 2021). En el Ecuador, el cierre de escuelas resultó en el incremento de la deserción escolar de los niños, niñas y adolescentes de todos los niveles educativos, especialmente en el nivel de bachillerato. El país fue uno de los veintidós de la región en adaptar una metodología de aprendizaje remoto para continuar con el servicio educativo en modalidades a distancia, con varias alternativas: en línea mediante plataformas web, por televisión, radio, o canales de comunicación en línea como WhatsApp, medios impresos, entre otros[1].

El sistema educativo ecuatoriano comprende los niveles: educación inicial, educación general básica (EGB), bachillerato general y educación superior y la mayor cantidad de estudiantes se concentra en el nivel EGB. Sin embargo, según varios informes del MINEDUC, se confirma que la deserción escolar se produce en mayor cantidad en el nivel de bachillerato. El nivel educativo bachillerato general en Ecuador comprende tres años de educación obligatoria a continuación de la Educación General Básica, y desde el 2011 entró en vigor el bachillerato general unificado.

[1] Esta información se detalla en el UPDATE #21 elaborado por el Fondo de las Naciones Unidas para la Infancia, UNICEF, publicado el 8 de febrero de 2021, donde se muestra un panorama de la situación de las escuelas en el contexto de la pandemia por Covid-19. En dicho informe, se detalla que 14 de los 36 países y territorios de la región han abierto sus escuelas parcialmente, entre ellos el Ecuador.

La deserción escolar se ha incrementado durante los últimos años en el país, pero se vio afectada especialmente por la pandemia de Covid-19, junto con los movimientos migratorios y las condiciones de pobreza, desigualdad e inseguridad. Varios autores (Loaiza, Romero, Ronquillo, García, & Diaz, 2023) señalan que la deserción escolar no es causada por un solo factor, sino por una combinación de factores interrelacionados. El contexto social, familiar e individual son responsables de la deserción escolar. Es decir, para analizar este fenómeno social hay que analizar factores dentro y fuera del contexto escolar. Otros autores afirman que:

> "*Los estudios disponibles sobre la deserción escolar en el Ecuador señalan varias causas y factores que contribuyen a este problema. Algunos de los hallazgos más comunes son los siguientes: factores socioeconómicos; desinterés y falta de motivación, problemas familiares; dificultades de salud y bienestar y problemas propios del sistema educativo ecuatoriano"* (Loaiza, Romero, Ronquillo, García, & Diaz, 2023).

De acuerdo con el informe Estadística Educativa del Ministerio de Educación de Ecuador (Ministerio de Educación de Ecuador, 2021) en el año lectivo 2019-2020 la tasa de abandono escolar nacional fue de 1,73% mientras que el siguiente periodo escolar 2020-2021 se incrementó a 1,77%, teniendo a su vez un menor número estudiantes inscritos a nivel nacional en comparación con el año escolar anterior. Para el periodo 2021-2022, la tasa de abandono continúo aumentando hasta alcanzar el 2,11% de estudiantes que abandonaron el sistema educativo. Estas cifras, sin contar con los niños, niñas y adolescentes no registrados en el sistema educativo o que nunca han asistido a un establecimiento escolar son preocupantes, pues reflejan la grave crisis educativa que se vio mayormente afectada por la pandemia y por las brechas sociales y económicas.

El desistir en la continuidad de los aprendizajes es un problema social grave que se ha transformado en un tema relevante por analizar y poder determinar las razones de este fenómeno en el contexto del país. En la revisión de la literatura, se identificó que las técnicas desarrolladas en minería de datos y en ciencia de datos pueden ser de gran utilidad para el desarrollo y elaboración de modelos de deserción estudiantil.

Con la idea de investigación se pretende elaborar un modelo predictivo de los principales factores de deserción escolar de los estudiantes de bachillerato en el Ecuador de los años 2012 al 2022, para sugerir medidas de prevención y fomentar la permanencia escolar en la educación.

1.4. Justificación

La deserción escolar es un problema medular de los sistemas educativos de América Latina desde hace muchos años atrás, que se vio exacerbado por la pandemia de Covid-19. Desde

febrero de 2020, la tasa de asistencia a clases se vio afectada fuertemente por la pandemia y el Ecuador, como el resto de los países de América Latina se acogió a la medida de cierre de escuelas para salvaguardar la salud de los estudiantes. Sin embargo, esta situación empeoró otras condiciones a las que se expusieron los niños y niñas, sobre todo los adolescentes; la pobreza, la desigualdad y los movimientos migratorios, provocaron un incremento en la deserción escolar en bachillerato, disminuyendo la tasa de finalización de estudios en todo el Ecuador.

De acuerdo con cifras del informe La educación en Ecuador Resultados educativos 2017-2018 del INEVAL (Instituto Nacional de Evaluación Educativa, 2018) la deserción escolar a nivel nacional se concentra mayormente en los niveles de bachillerato que en los demás niveles (Educación General Básica). Esta tendencia se ha mantenido durante los consecuentes periodos escolares. A pesar de los esfuerzos del Ministerio de Educación con la implementación de programas como el Plan Educativo: aprendamos juntos en casa, o la iniciativa "Todos al Aula" para responder a las necesidades educativas de los adolescentes y reinsertarlos al sistema educativo formal, la deserción escolar continúa presentándose hasta el día de hoy, afectando las vidas de los niños, niñas y adolescentes.

El Ministerio de Educación ha desarrollado diversas metodologías para cuantificar la deserción escolar, con indicadores como la tasa de abandono escolar. La tasa de abandono escolar se define como el "número de estudiantes contabilizados al final de un período escolar que abandonan un determinado grado o curso de estudios, expresado como porcentaje del total de estudiantes matriculados al final del mismo grado o curso de estudios y periodo escolar" (Ministerio de Educación de Ecuador, 2021) y sirven para analizar estas cifras y determinar los desafíos que conlleva la respuesta educativa para aumentar la retención de estudiantes en el sistema educativo, y garantizar la continuidad de sus aprendizajes, por ende, la finalización de sus estudios.

Esta investigación pretende determinar cuáles son los factores de deserción escolar de los estudiantes de bachillerato en el Ecuador mediante un modelo predictivo, para así poder evaluar los resultados obtenidos, y en función de ello, sugerir medidas de prevención de la deserción y fomentar la permanencia de los adolescentes en el sistema educativo, así como identificar las principales razones de su deserción. La educación es un derecho fundamental de vital importancia para el desarrollo individual y social de la adolescencia, sin dejar de mencionar que es una herramienta para mejorar las condiciones de vida y laborales en el largo plazo. Además, el estudio de los niveles de retención escolar es sumamente importante para los países y los gobiernos, ya que implica encaminar el correcto desarrollo del capital humano, pues saber qué decisiones de política pública tomar para mejorar las condiciones del sistema educativo puede llevar a mejores niveles de vida de la población y, por ende, mayor generación de recursos económicos para el país.

Para determinar los factores de la deserción escolar en las instituciones educativas del Ecuador, se ha optado por la técnica de modelos predictivos dentro de la minería de datos, ya que según la investigación realizada en varios artículos científicos, las técnicas estadísticas tradicionales no son suficientes para analizar grandes volúmenes de datos, por ello, se hará uso de una de las técnicas de aprendizaje supervisados más utilizadas en minería de datos mediante la generación de algoritmos para diseñar modelos predictivos.

Varios autores que han llevado a cabo estudios similares en el campo de la minería de datos educativa (EDM, por sus siglas en inglés de Educational Data Mining) (Cornejo Sifuente, y otros, 2023) (Garača & Čukušić, 2010), (Hernández González, y otros, 2016) recomiendan esta metodología por su precisión en los resultados a la hora de hallar a los posibles estudiantes desertores y también para hallar un perfil de los estudiantes con las mismas probabilidades de deserción. De igual manera, los árboles de decisión son una técnica de aprendizaje supervisado utilizada comúnmente para determinar patrones de deserción escolar en el ámbito educativo según la revisión de la literatura (Cuji, Gavilanes, & Sánchez, 2017), (Hernández González, y otros, 2016) (Merchan & Duarte García, 2016) y es la segunda metodología que se propone en caso de que la información no cumpla con los requerimientos para llevar a cabo la técnica de regresión logística.

1.5. Objetivos

1.5.1. Objetivo general

Elaborar un modelo predictivo de los factores de deserción escolar de las instituciones educativas de bachillerato en el Ecuador de los años 2012 – 2022, mediante el uso de técnicas de minería de datos.

1.5.2. Objetivos específicos

- Revisar la literatura relacionada a los modelos predictivos de deserción escolar mediante el uso de minería de datos, para identificar metodologías y herramientas en investigaciones similares.
- Identificar los factores asociados con la deserción escolar en el nivel de secundaria bachillerato en Ecuador utilizando herramientas de minería de datos, para determinar la relación entre las variables disponibles de las instituciones educativas y la deserción escolar.
- Construir un modelo predictivo de los factores de deserción escolar que permita identificar patrones y factores asociados con la deserción escolar a nivel de institución educativa, utilizando técnicas de minería de datos.
- Validar y evaluar el modelo predictivo construido según la utilidad y confiabilidad de los hallazgos.

2. CAPÍTULO II

2.1. Marco teórico y conceptual

2.1.1. Educación

La educación es un derecho humano fundamental, habilitante de otros derechos como lo son la nutrición, la salud, acceso a agua segura, protección de la infancia, entre otros. Además, las escuelas son un lugar seguro para los niños, niñas y adolescentes que reciben clases. Sin embargo, las complejidades de las esferas económicas, sociales, ambientales, políticas afectan rotundamente la continuidad de la trayectoria educativa de los niños y niñas. El artículo 26 de la Declaración Universal de los Derechos Humanos establece que "toda persona tiene derecho a la educación. La educación debe ser libre, al menos para los niveles elementales y fundamentales" (United Nations, 1948).

Uno de los graves problemas que afectó a la continuidad de los estudios es la pandemia de Covid-19, situación que cambió radicalmente la forma de enseñanza en todo el globo. Más de 190 países cerraron masivamente las escuelas en todo el mundo cuando la Organización Mundial de la Salud declaró este nuevo virus una pandemia[2]; asimismo, los países y territorios de la región tuvieron que acogerse a esta medida de prevención, y cerrar todas las escuelas hasta nuevo aviso, para salvaguardar la salud de los niños, niñas, adolescentes, docentes y otro personal educativo.

Este hecho provocó la búsqueda de nuevas alternativas a la educación presencial, por lo que se adoptaron nuevas medidas para continuar brindando los servicios educativos. Varios países crearon plataformas web dedicadas sólo a las clases en modalidad virtual, con recursos pedagógicos enfocados en los diferentes niveles educativos, muchos optaron por transmitir las clases mediante medios comunicacionales como la televisión, mientras que otros lo hicieron mediante la radio.

Según la Comisión Económica para América Latina y el Caribe (CEPAL):

> *"Gran parte de las medidas que los países de la región han adoptado ante la crisis se relacionan con la suspensión de las clases presenciales en todos los niveles, lo que ha dado origen a tres campos de acción principales: el despliegue de modalidades de aprendizaje a distancia, mediante la utilización de una diversidad de formatos y plataformas (con o sin uso de tecnología); el apoyo y la movilización del personal y las comunidades educativas, y la atención a la salud y el bienestar integral de las y los estudiantes"* (Comisión Económica para América Latina y el Caribe CEPAL, 2020).

Sin embargo, a pesar de los esfuerzos de los países por mantener la educación, muchos de los adolescentes en situación de vulnerabilidad, de bajos recursos, o en situación de movilidad

[2] Según datos de la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO) a mediados de mayo de 2020, más de 1.200 millones de estudiantes de todos los niveles de enseñanza, en todo el mundo, habían dejado de tener clases presenciales en la escuela.

humana no pudieron continuar sus estudios por distintas barreras, como la falta de recursos económicos, la falta de dispositivos tecnológicos para asistir a clases, el trabajo infantil, la desigualdad, entre otros factores. Sin dejar de mencionar que el deterioro de los sistemas educativos en la región ya se encontraba en decadencia incluso antes de la pandemia por las desigualdades sociales, pobreza y pobreza extrema.

Considerando esta grave afectación al sistema educativo nacional, es necesario recalcar que en 2021 alrededor de 90.000 estudiantes ya se encontraban fuera del sistema educativo y lograron ser identificados por el Ministerio de Educación, consecuentemente la pandemia agravó de sobremanera esta situación (Fondo de las Naciones Unidas para la Infancia, UNICEF, 2021).
De acuerdo con UNICEF, no asistir presencialmente a las escuelas afecta su bienestar, seguridad y desarrollo, pues es allí donde reciben educación, alimento y protección, donde juegan, hacen amigos y reciben el apoyo de sus docentes (Fondo de las Naciones Unidas para la Infancia, UNICEF, 2021).

2.1.2. Deserción escolar

De acuerdo con (Pachay & Rodríguez, 2021) la deserción escolar es el abandono de las actividades académicas de un individuo, que por diversas situaciones como económicas, políticas, sociales, familiares, ambientales o de salud, ocurre cuando las personas dejan atrás el proceso de educación o formación.

La deserción escolar en el Ecuador es calculada mediante un indicador conocido como "Tasa de abandono escolar" y se define como el "número de estudiantes contabilizados al final de un periodo escolar que abandonan un determinado grado o curso de estudios, expresado como porcentaje del total de estudiantes matriculados al final del mismo grado o curso de estudios y periodo escolar"

Las variables que intervienen en el cálculo de la tasa de abandono escolar son:

- Periodo escolar: Año académico establecido en meses, en el que los estudiantes reciben enseñanza en un establecimiento educativo.
- Matrícula: Registro o inscripción de los estudiantes que van a realizar sus estudios en un grado o curso en período dado, dentro de un centro de enseñanza.
- Descomposición de la matrícula: Situación de la matrícula del estudiante en el período escolar en el que se encuentra actualmente. Se compone de los siguientes estados: promovido, no promovido, abandono (Ministerio de Educación de Ecuador, 2021).

2.1.3. Factores que provocan la deserción escolar

Como se ha visto anteriormente, la deserción escolar no es causada por un solo factor, sino que es una combinación de factores que están interrelacionados. En el Ecuador, son diversas las causas por las que no se culminan los estudios y exista un alto abandono de los estudios por

parte de los estudiantes, sobre todo en las instituciones educativas públicas. Estas causas están explicadas por factores socioeconómicos, políticos y culturales. Algunos de los factores que provocan la deserción escolar en Ecuador son:

- Apoyo limitado a los estudiantes.
- Falta de recursos y/o infraestructura en las escuelas.
- Poca relevancia de la educación.
- Factores individuales que afectan al estudiante, como son: falta de interés, malos comportamientos y actitudes, bajo rendimiento académico, o repetición de grados académicos.
- Pobreza, pobreza extrema y factores económicos.
- Trabajo infantil o a edad temprana.
- Discapacidad o multi discapacidad.
- Embarazo adolescente.
- Factores que ponen en riesgo la integridad del adolescente como la violencia, el desplazamiento o situación de movilidad humana, entre otras (Loaiza, Romero, Ronquillo, García, & Diaz, 2023).

Según varios estudios realizados en la región, los niños se encuentran más vulnerables que las niñas durante su trayectoria escolar. Un niño tiene menos probabilidades de terminar sus estudios, según datos de la CEPAL:

> "*Aunque globalmente las niñas tienen menos posibilidades que los niños de matricularse en la escuela, y muchas todavía son excluidas de la educación, los niños corren más riesgo de abandonar la escuela prematuramente (IEU, 2019), en particular aquellos que viven en la pobreza. Es preciso desarrollar estrategias para prevenir la desvinculación y el abandono escolar de los niños, así como evaluaciones de dichas estrategias para establecer lo que funciona y lo que no*" (Comisión Económica para América Latina y el Caribe CEPAL, 2020).

Según (Forbes Ecuador, 2022) en 2022, el 4.1% de niños y adolescentes abandonó la escuela o colegio y existió una disminución de un 3% de estudiantes matriculados en ese año, así lo indican las cifras del Ministerio de Educación y el Instituto Nacional de Estadística y Censo (INEC).

De acuerdo con el informe del Ministerio de Educación del Ecuador muestra que la tasa de deserción escolar en secundaria es del 9,7%, lo que significa que cerca de 1 de cada 10 estudiantes de secundaria abandona la escuela antes de finalizar sus estudios (Ministerio de Educación de Ecuador, 2020).

El Ministerio de Educación reportó para el año lectivo 2023 la cifra más baja de matriculación escolar en los últimos seis periodos académicos identificándose el trabajo como una de las

principales causas por las que los estudiantes abandonan las aulas de clases (Zambrano Ostaiza & Guaña Bravo, 2024).

La ilustración 1 muestra los factores que provocan la deserción escolar en Ecuador.

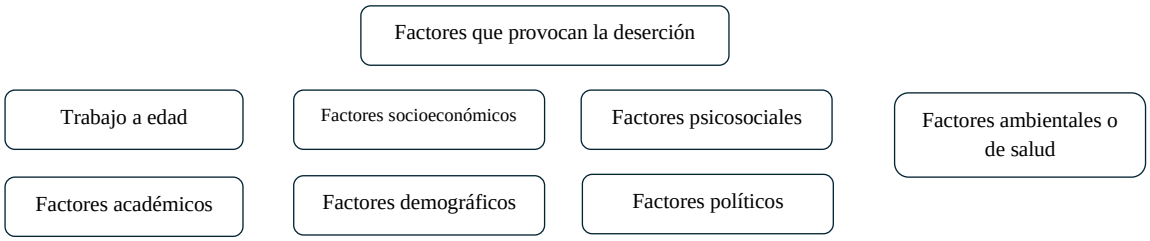


Ilustración 1 Factores que provocan deserción escolar (Loaiza, Romero, Ronquillo, García, & Diaz, 2023, adaptado por (Melo 2024)

Para responder a las necesidades educativas, el país desarrolló varios planes a ser aplicados a nivel nacional, con el fin de frenar la deserción escolar. En el contexto de la emergencia sanitaria por Covid-19, la mayoría niños, niñas y adolescentes dejaron de ir a la escuela por motivos de seguridad y salvaguarda, por lo que el Ministerio de Educación del Ecuador (MINEDUC) elaboró la estrategia de respuesta educativa para garantizar la continuidad de los estudios de los estudiantes del sistema educativo ecuatoriano. La entidad firmó el Decreto Ejecutivo de Políticas Educativas, con el cual se trabajará bajo cinco ejes fundamentales para garantizar que la oferta educativa llegue a todos los niños, niñas y adolescentes.

- Primer Eje “Encontrémonos”: Se centrará en las brigadas nacionales de docentes, quienes interactuarán con los estudiantes para evaluar su situación en distintos "Puntos de Encuentro" a lo largo del país. Estos puntos servirán como espacios de conexión entre la escuela, las familias y la comunidad.
- Segundo Eje “Todos”: Buscará fortalecer la escuela rural y comunitaria, posicionándola como un motor de desarrollo para las comunidades. De las 1.500 instituciones educativas que fueron cerradas, hasta 2021 se reabrieron 141, y el objetivo es reabrir 900 más durante el presente gobierno.
- Tercer Eje “Libres y Flexibles”: Promoverá una mayor autonomía y libertad en las instituciones educativas. Se impulsará que las 28 editoriales encargadas de la producción de recursos pedagógicos, como libros de texto, tengan más independencia y responsabilidad sobre los contenidos y la calidad de más de 174 textos escolares.
- Cuarto Eje “Fuertes”: Se enfocará en mejorar el bienestar de los maestros y en flexibilizar la educación. Se implementará un nuevo escalafón docente con el objetivo de dignificar la profesión.
- Quinto Eje “Excelencia Educativa”: Se dedicará a mejorar la calidad educativa tanto en el ámbito pedagógico como tecnológico. En los primeros 60 días, se entregarán 750 tabletas con el apoyo del sector privado. (Ministerio de Educación de Ecuador MINEDUC, 2021).

Una de las estrategias que se desarrolló en el marco de estos cinco ejes es el Plan Educativo Covid-19 para estudiantes de tercer año de bachillerato, el cual cuenta con varias dimensiones para priorizar los conocimientos básicos, herramientas pedagógicas y metodológicas, psicoemocionales y psicosociales, para ser desarrollado desde el hogar. Una de las metodologías que implica es el Aprendizaje por Proyectos.

Asimismo, se desarrollaron otros planes educativos como el Plan "Aprendemos Juntos en Casa" el cual contiene varios recursos educativos tanto para estudiantes como para docentes. Otra estrategia implementada es el diseño de un currículo para la emergencia[3], con la colaboración de la UNESCO junto al MINEDUC, este último señala que se elaboró esta herramienta para priorizar los objetivos de aprendizaje y el desarrollo de contenidos esenciales para que los estudiantes puedan acceder al siguiente año de estudios (Ministerio de Educación, 2020).

Dentro del eje uno "Encontrémonos" orientado a la reactivación de las instituciones educativas y el desarrollo de planes de reinserción escolar y nivelación de estudiantes en edad escolar que se encuentran fuera del SNE, esta cartera de estado ha venido implementando la estrategia "Todos al Aula" desde noviembre de 2022 (Ministerio de Educación de Ecuador, 2021). Todos al Aula es una estrategia implementada por el MINEDUC apoyada por la UNESCO, que tiene el objetivo de garantizar el acceso a la educación y la reinserción de niños, niñas y adolescentes entre los 5 y 14 años, dirigida también a aquellos que se encuentran fuera del sistema educativo.

La razón más común en el Ecuador es la falta de recursos económicos, la pobreza y la pobreza extrema, sobre todo en zonas rurales del país en la región Costa y Sierra. Varias familias no cuentan con los dispositivos tecnológicos adecuados para que los niños y niñas continúen recibiendo clases en las modalidades virtuales propuestas, así como no cuentan con los servicios de internet y señal satelital necesaria para conectarse a las plataformas virtuales. De acuerdo con cifras de UNICEF:

> "*En Ecuador, solo el 37 por ciento de los hogares tiene acceso a internet, lo que significa que 6 de cada 10 niños no pueden continuar sus estudios a través de plataformas digitales. La situación es más grave para los niños de zonas rurales, solo el 16 por ciento de los hogares tiene este servicio*" (Fondo de las Naciones Unidas para la Infancia, 2020).

[3] El currículo para la emergencia fue una iniciativa desarrollada durante la pandemia, en respuesta a la emergencia sanitaria por Covid-19. El diseño de la herramienta fue contextualizado a la realidad ecuatoriana y a los desafíos que planteaba la emergencia en ese entonces.

2.1.5. Proceso de descubrimiento del conocimiento

El descubrimiento de conocimiento en bases de datos (*Knowledge Discovery Databases*, por sus siglas en inglés) es un proceso automático en el que se combinan descubrimiento y análisis. El proceso consiste en extraer patrones en forma de reglas o funciones, a partir de los datos, para que el usuario los analice (Timarán-Pereira, Hernández-Arteaga, Caicedo-Zambrano, Hidalgo-Troya, & Alvarado-Pérez, 2016).

Este proceso implica seguir una serie de pasos de selección de datos, preprocesamiento, transformación, minería de datos y la interpretación, para finalmente generar conocimiento. Sin embargo, este proceso no es lineal, se puede volver a las fases anteriores para mejorar los resultados de este proceso.

La minería de datos o *data mining* se define como un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos (Pérez & Santín, 2008). Además, en la actualidad es ampliamente utilizada en diversos campos como la ciencia, los negocios, la salud, la banca, las telecomunicaciones, la detección de fraudes y la educación.

La minería de datos en la educación ha surgido debido a la creciente accesibilidad de los datos educativos y, por lo tanto, a la necesidad de analizar esta cantidad masiva de datos (Kumar, Singh, & Handa, 2017). Este campo de investigación multidisciplinar es usado para analizar datos educativos mediante técnicas de minería de datos, especialmente cuando se quiere comprobar el rendimiento estudiantil en un futuro próximo con su historial educativo previo (Kumar, Singh, & Handa, 2017).

La minería de datos en la educación (*Educational Data Mining*) ya se aplica en varios ámbitos para predecir el rendimiento académico estudiantil, en sistemas de gestión para el aprendizaje de contenidos, en los sistemas inteligentes de aprendizaje, entre otros usos (Cornejo Sifuente, y otros, 2023).

2.1.6. Técnicas de minería de datos

Como se ha mencionado anteriormente, existen varias ciencias que hacen uso de técnicas de minería de datos, como la agricultura, la medicina, las telecomunicaciones, inclusive, los negocios para facilitar la toma de decisiones de la empresa, y en el ámbito educativo también es ampliamente utilizado para ofrecer mejores servicios educativos, mejorar las condiciones del servicio, etc. Las técnicas de minería de datos persiguen el descubrimiento automático del conocimiento contenido en la información almacenada de modo ordenado en grandes bases de datos (Pérez & Santín, 2008).

Las técnicas de minería de datos se clasifican principalmente en técnicas predictivas, donde las variables pueden ser clasificadas como variables de resultado y predictivas, y las técnicas

descriptivas, en donde inicialmente todas las variables tienen el mismo estatus (Pérez & Santín, 2008).

Las técnicas predictivas especifican el modelo para los datos en base a un conocimiento teórico previo. El modelo supuesto para los datos debe contrastarse después del proceso de minería de datos antes de aceptarlo como válido. Podemos incluir entre estas técnicas todos los tipos de regresión, series temporales, análisis de la varianza y covarianza, análisis discriminante, árboles de decisión y redes neuronales. Pero, tanto los árboles de decisión, como las redes neuronales y el análisis discriminante son a su vez técnicas de clasificación que pueden extraer diferentes perfiles de comportamiento o clases, siendo el objetivo construir un modelo que permita clasificar cualquier nuevo dato (Pérez & Santín, 2008).
Según el objetivo del análisis de datos, los algoritmos utilizados se clasifican en dos grandes grupos: supervisados y no supervisados (Theodoridis & Koutroumbas, 2006).

En las técnicas descriptivas no se supone la existencia de variables dependientes ni independientes y tampoco se supone la existencia de un modelo previo para los datos. En este grupo se incluyen las técnicas de agrupación o *clustering* y segmentación (que también son técnicas de clasificación en cierto modo), las técnicas de asociación y dependencia, las técnicas de análisis exploratorio de datos y las técnicas de reducción de la dimensión (factorial, componentes principales, correspondencias, etc.) (Pérez & Santín, 2008).

La ilustración 2 muestra la clasificación de las técnicas de minería de datos.

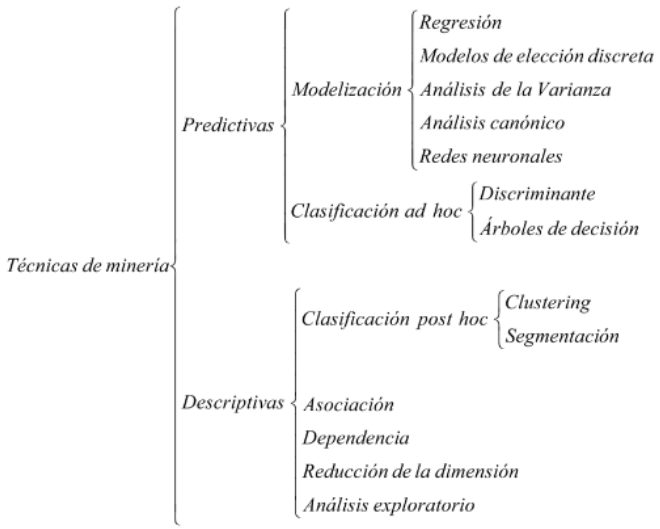


Ilustración 2 Técnicas de minería de datos (Pérez & Santín, 2008), adaptado por (Melo 2024)

El uso de conjuntos de datos etiquetados para entrenar algoritmos que clasifiquen datos o predigan resultados de forma precisa se conoce como aprendizaje supervisado, también conocido como *machine learning* supervisado, es una subcategoría del *machine learning* y la inteligencia artificial (IBM).

El aprendizaje supervisado puede clasificarse en dos tipos de problemas durante la minería de datos:

- La clasificación: la clasificación utiliza un algoritmo para asignar con precisión datos de prueba en categorías específicas.
- La regresión: la regresión se utiliza para comprender la relación entre variables dependientes e independientes.

2.1.7. Algoritmos de aprendizaje supervisado

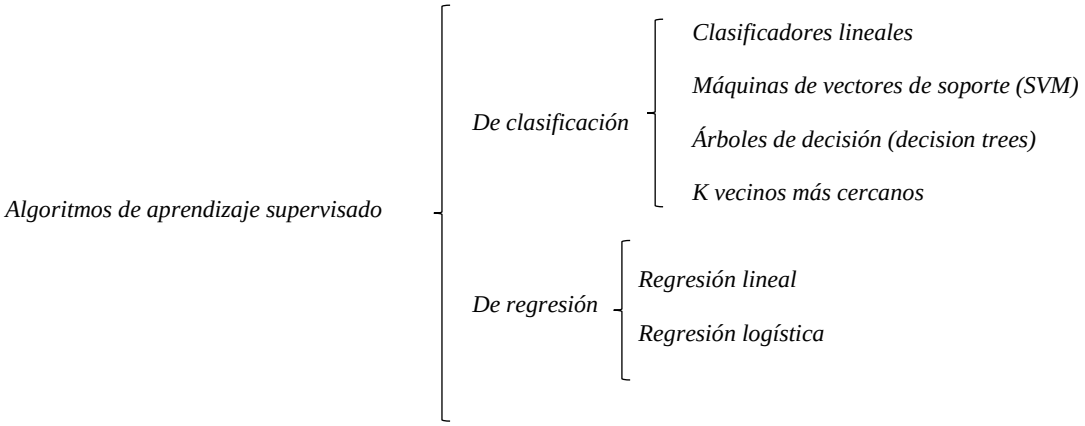


Ilustración 3 Algoritmos de aprendizaje supervisado (Pérez & Santín, 2008), adaptado por (Melo 2024)

El objetivo de los algoritmos de aprendizaje supervisado o inductivo es encontrar la predicción dadas las entradas; extrae patrones de ejemplos conocidos y usa esta información extraída para diseñar un resultado replicable con un nuevo set de datos. La característica principal de este algoritmo es que los datos están etiquetados. Se utilizan varios algoritmos y técnicas de cálculo en los procesos de *machine learning* supervisados.

La regresión logística es una técnica estadística ya muy bien conocida que se usa para modelar los resultados de encuestas. Existen varias implementaciones de regresión logística en la investigación estadística, que utilizan diferentes técnicas de aprendizaje (Hernández González, y otros, 2016).

2.1.8. Modelos para predecir la deserción escolar

La literatura ha mostrado que principalmente las causas de la deserción escolar son socioeconómicas, sin embargo, todavía hay una falta de mecanismos que permitan recolectar los datos y analizarlos de manera más precisa, ya que usualmente los métodos usados para identificar los niveles de deserción escolar son mediante estadísticas y encuestas. Variables como el rendimiento académico, asistencia, situación socioeconómica y factores demográficos han sido considerados como indicadores tradicionales de riesgo de deserción. Por ello, es importante profundizar en los métodos para predecir la deserción escolar en el Ecuador.

En la última década, la ciencia de datos y el uso de grandes cantidades de datos a dado paso a la aplicación técnica de minería de datos en el ámbito educativo, para identificar los factores o variables predictoras de la deserción escolar. Esta identificación se puede realizar a través de diferentes metodologías, como el uso de la econometría, estadística y modelos de minería de datos.

Existen varios modelos de minería de datos para predecir o estimar la deserción escolar. Los modelos predictivos de clasificación y regresión son las metodologías más utilizadas para desarrollar modelos más precisos. En un estudio desarrollado por (Manandhar & Sthapit, 2005) se utilizó un modelo de regresión logística para predecir la deserción escolar de los niños y niñas en una escuela del distrito de Chitwan, en Nepal, en el cual se analizó variables sociales relacionadas con el entorno escolar de los niños y niñas para predecir el abandono escolar, como la edad, casta, trabajo en casa y niño o niña no interesado en el estudio, demostrando que la deserción escolar está explicada en estas variables.

La publicación "La iniciación laboral temprana y la deserción escolar en los estudiantes de bachillerato" (Zambrano Ostaiza & Guaña Bravo, 2024) realizó una investigación en la Unidad Educativa Carmelina Teófila Moreira sobre los factores determinantes de la deserción escolar en dicha institución. La investigación se centra en el nivel de bachillerato, pues es donde es más recurrente la iniciación laboral temprana, lo que provoca el abandono de la trayectoria educativa antes de culminar la educación secundaria. La técnica para recopilar la información fue la encuesta realizada a varios docentes y estudiantes de la unidad educativa, participaron 222 personas y los resultados arrojaron que el factor principal para la iniciación laboral temprana y en consecuencia abandonar los estudios de bachillerato son las necesidades económicas de las familias, ya que los adolescentes se sienten presionados por esta necesidad y deciden buscar un medio de subsistencia.

Otro de los hallazgos en materia de deserción escolar es el artículo "Identificación de los factores de la deserción académica en el sistema educativo del Ecuador" (Loaiza, Romero, Ronquillo, García, & Diaz, 2023) el cual realizó un estudio cuantitativo de los factores que inciden en la deserción escolar del sistema educativo ecuatoriano, con un diseño no experimental de corte descriptivo cuantitativo. El estudio utilizó datos del Instituto Nacional de Estadística y Censos (INEC) del periodo 2010 al 2021. Los resultados muestran que a pesar de que la deserción escolar ha mantenido una tendencia a lo largo del periodo estudiado, aumentó en el periodo 2010-2012, además, el sexo masculino predomina en cuanto a deserción de los estudios a nivel nacional. En cuanto a los niveles educativos, en bachillerato la tasa de abandono escolar fue de 3,25% para el 2021, en comparación a EGB donde la tasa de abandono fue de 1,41%. El estudio recomienda analizar las causas subyacentes a la deserción escolar por nivel educativo para abordarlas de manera eficaz. Finalmente, el estudio muestra que una de las causas por las que los estudiantes abandonan sus estudios es el nivel económico, seguido

del bajo rendimiento académico, la falta de acceso a la educación, los factores familiares y finalmente los problemas de salud que presenta el estudiante.

El artículo "Modelo Predictivo de la Deserción Escolar en Educación Superior: una Aproximación desde la Minería de Datos Utilizando la Metodología CRISP-DM" (Cornejo Sifuente, y otros, 2023) elaboró un modelo para predecir los casos de estudiantes de educación superior en riesgo de deserción escolar mediante el uso de técnicas de minería de datos. El estudio contó con la participación de 1.374 estudiantes y se analizaron factores familiares, individuales, académicos, sociales y de compromiso institucional. La metodología utilizada fue CRISP-DM y se usaron algoritmos de clasificación basados en programación genética, comparándolos con otros algoritmos de clasificación como árboles de decisión. Los resultados muestran que el modelo de clasificación que mejor predijo los factores fue Ramdom Tree, y los factores que influyen en la deserción escolar son el semestre que cursa el alumno, si el alumno presentó exámenes extraordinarios o no y su región de procedencia, la edad y el semestre, el promedio de bachillerato, la calificación obtenida en su examen de admisión y si índice CENEVAL. El estudio demostró la importancia de tomar medidas para prevenir la deserción escolar desde la institución.

El estudio "Hacia la construcción de un modelo predictivo de deserción académica basado en técnicas de minería de datos" (Sotomonte Castro, Rodríguez Rodríguez, Montenegro Marín, Gaona García, & Castellanos, 2016) hace uso de técnicas de minería de datos para determinar las causas de la deserción escolar en la educación superior de Colombia. Se utilizó la metodología CRISP-DM y un modelo de árbol de decisión con el uso de la herramienta WEKA. Los datos utilizados fueron los récords académicos de los estudiantes de la facultad de ingeniería de la Universidad Distrital Francisco José de Caldas de los años 2009 y 2015. Los resultados del estudio muestran que uno de los factores determinantes para abandonar los estudios es el número de materias que tiene el estudiante y el nivel socio económico.

El artículo "Un Modelo Predictivo de Deserción Escolar para la República Dominicana" (Llaugel & González-Disla, 2016) se basó en la construcción de un modelo predictivo de minería de datos para determinar los patrones de deserción escolar de los estudiantes de nivel básico y medio en el sistema educativo dominicano. El estudio se desarrolló con información de 72 centros educativos del distrito escolar Los Alcarrizos en los periodos del 2009 al 2014. Se usó la metodología CRISP-DM para desarrollar el modelo, y se usaron las herramientas IBM SPSS y RapidMiner para el análisis de datos. Los atributos considerados para elaborar el modelo predictivo fueron la condición académica del alumno, el tiempo de permanencia del alumno en el sistema educativo al momento de pausar sus estudios y antes de pasar a condición de deserción, el último grado alcanzado, la cantidad de abandonos previos, tiempo de deserción transcurrido, cantidad de reprobaciones, cantidad de promociones, si se ha transferido de centro educativo antes de pasar a condición de deserción o egresado. El modelo da un resultado óptimo

sobre riesgo de deserción escolara para la muestra usada de 20.000 estudiantes en los 72 centros educativos de Los Alcarrizos.

Otro artículo sobre predicción de la deserción escolar "Predicting School Failure Using Data Mining" (Márquez, Ventura, & Romero, 2011) utilizó datos reales de 670 estudiantes de nivel escolar medio en Zacatecas, México con 77 atributos obtenidos de registros académicos y una encuesta elaborada a los estudiantes. Mediante el uso de la herramienta WEKA se desarrollaron cinco algoritmos de inducción (JRip, NNge, OneR, Prism y Ridor) y se usaron también árboles de decisión (J48, SimpleCart, ADTree, RandomTree y REPTree). El estudio concluyó en que el mejor método para predecir el riesgo de deserción escolar es ADTree y luego se realizó un ajuste en los atributos para seleccionar los 15 más relevantes para el modelo sin perder el rendimiento en el modelo de clasificación. Los atributos que explican la deserción en este estudio son las calificaciones en Física, Humanidades, Matemáticas e Inglés; otros factores son la sobre edad escolar (mayor de 15 años de edad), tener más de un hermano/hermana, asistir a clases vespertinas y la baja motivación para estudiar.

Dentro de la revisión de la literatura se encontró artículos que usaban modelos de regresión logística, como el artículo "Modelo de regresión logística para la estimación de la deserción escolar del Posgrado en la Universidad Técnica de Manabí, Ecuador" (Solís Ventura, Quiroz Fernández, & Fosado Téllez, 2022) el cual presenta un modelo de regresión logística binaria para la predicción de la deserción escolar en el nivel de posgrado. Los datos fueron obtenidos del sistema de gestión académica con una muestra de 729 estudiantes en estudios de posgrado. Los resultados indicaron que los factores asociados a la deserción escolar son el estado civil, los ingresos percibidos, la situación laboral y la edad promedio del estudiante.

El artículo "Prediction of university dropout through technological factors: a case study in Ecuador" (Alban Taipe & Sánchez, 2018) se enfoca en determinar los factores tecnológicos que influyen en la deserción escolar en el nivel universitario. Utiliza un enfoque de aprendizaje automático basado en regresión logística y árboles de decisión, los atributos usados para determinar el modelo son factores como adicción al internet, adicción a las redes sociales y adicción a la tecnología. El estudio dio como resultado que el factor que más influye en la deserción escolar es la adicción al internet, que corresponde al uso de internet sin fines académicos.

Otro de los modelos desarrollados usando árboles de decisión se presenta en el artículo "Modelo predictivo de deserción estudiantil basado en arboles de decisión" (Cuji, Gavilanes, & Sánchez, 2017) donde se busca predecir la probabilidad de que un estudiante universitario abandone sus estudios considerando atributos como rendimiento académico y variables del entorno personal. Se usaron técnicas de clasificación basadas en árboles de decisión junto al algoritmo CART para contar con variables nominales y cuantitativas. El árbol de decisión fue

el algoritmo que mejor predijo la probabilidad de desertar, siendo las variables más importantes el nivel educativo y notas académicas.

3. CAPÍTULO III

3.1. Metodología

La metodología de la presente investigación se basará en el proceso de la minería de datos Descubrimiento de Conocimiento en Bases de Datos (*Knowledge Discovery in Databases KDD*, por sus siglas en inglés). Según (Nigro, Xodo, Corti, & Terren) el Descubrimiento de Conocimiento en Bases de Datos se define como:

> "*El descubrimiento de conocimiento en bases de datos es un campo de la inteligencia artificial de rápido crecimiento, que combina técnicas del aprendizaje de máquina, reconocimiento de patrones, estadística, bases de datos, y visualización para automáticamente extraer conocimiento (o información), de un nivel bajo de datos (bases de datos. El KDD es un área que está tomando importancia dado el crecimiento actual de las bases de datos (incluyendo bases de datos relacionales, bases de datos de objetos, bases de datos spatial-time y otras), y de la capacidad del hardware disponible para procesar estos datos*" (Nigro, Xodo, Corti, & Terren).

Se emplea para realizar la extracción automatizada de conocimiento a partir de grandes volúmenes de datos. Este proceso es de naturaleza iterativa, lo que permite su aplicación repetida tantas veces como sea necesario hasta obtener la información deseada. Esta metodología se compone de cinco etapas:

- Selección de los datos
- Preprocesamiento de los datos
- Transformación de los datos
- Minería de datos
- Interpretación de los resultados

3.2. Etapas del proceso de KDD

Como se mencionó anteriormente, el proceso de KDD es un proceso interactivo e iterativo, que involucra varias etapas en la toma de decisiones: selección, preprocesamiento o limpieza, transformación o reducción, minería de datos e interpretación. A continuación, se detalla cada una de las etapas para el proyecto de investigación:

3.2.1. Etapa de selección

La etapa de selección, según Fayyad (1996) consiste en que, una vez definidas las metas del proceso de KDD, desde la perspectiva del usuario final, se crea un conjunto de datos objetivo, de preferencia una muestra representativa del mismo.

3.2.2. Privacidad de los datos

La fuente de información es el portal de datos abiertos del MINEDUC, por lo que esta información es de acceso público y gratuito.

3.2.3. Etapa de preprocesamiento o limpieza

Según Fayaad (1996) en esta etapa se analiza la calidad de los datos, se aplican operaciones de eliminación de datos ruidosos o atípicos (también conocidos como outliers), se definen estrategias para manejar datos faltantes y técnicas para su reemplazo, así como el uso de gráficas de cajas o boxplot para una mejor comprensión del contexto del dato. Para el tratamiento de datos atípicos, como por ejemplo, sobreedad se deberá analizar la naturaleza de los datos y tomar una decisión en base a ello o determinar si se trata de un outlier contextual, para así garantizar la limpieza de los datos.

En esta etapa, además, se realizarán técnicas para imputación de datos faltantes si es el caso, como la aplicación de medidas de tendencia central (media, moda y mediana) o se optará por eliminar los registros de estudiantes con información incompleta (valores nulos). Se espera utilizar otras técnicas como la transformación de datos (normalización escalando los datos), integración de datos si así lo requiere.

3.2.4. Etapa de transformación y/o reducción

En esta etapa se buscará características útiles para representar los datos en función de los objetivos del proceso de KDD definidos inicialmente. En este caso, para un mejor entendimiento de los datos, se realizarán procesos de reducción de dimensiones como eliminación de atributos redundantes, tales como fecha de nacimiento y edad y se seleccionará solo los más relevantes de los datos disponibles. Se hará uso de técnicas de reducción de dimensiones como componentes principales en el caso de variables numéricas y correspondencias para el caso de variables categóricas (Han, Kamber, & Pei, 2001).

3.2.5. Etapa de minería de datos (data mining)

El objetivo de esta etapa es la búsqueda y descubrimiento de patrones de interés aplicando técnicas de clasificación como ya se ha mencionado anteriormente (Quinlan, 1986). Una vez que se ha culminado con la etapa de preprocesamiento de los datos, se procederá a realizar una exploración en profundidad de su correlación con la deserción escolar. Posteriormente, se aplica el sistema de minería de datos basado en el algoritmo de regresión logística para predecir la deserción escolar. El modelo predictivo de minería de datos se someterá a evaluación con datos de entrenamiento y de prueba. Se plantea utilizar el 80% del dataset para datos de entrenamiento y el 20% para la prueba del modelo.

3.2.6. Etapa de interpretación y evaluación de resultados

En esta etapa se interpretan los patrones descubiertos y en el caso de ser necesario, se retorna a las etapas anteriores de procesamiento de datos y minería de datos para posteriores iteraciones. Un clasificador de tipo binario se puede identificar por las clases: positiva y negativa.

- Clase positiva: Se refiere a la instancia que es clasificada como perteneciente a la categoría que el clasificador busca identificar.
- Clase negativa: Se refiere a la instancia que es clasificada como no perteneciente a la categoría que el clasificador intenta identificar.

En el caso de los algoritmos de aprendizaje supervisado binario, una de las técnicas más utilizadas para evaluar un algoritmo de regresión logística es mediante la matriz de confusión. La matriz de confusión muestra los valores reales de las entradas y los valores predichos mediante el algoritmo de regresión logística u otro algoritmo utilizado. La matriz muestra dos posibles valores, los positivos y los negativos, el análisis se realiza observando los valores predichos que se encuentran en la diagonal principal de la matriz de dos por dos.
La ilustración 4 muestra una matriz de confusión.

		Predicción	
		Positivo	**Negativo**
Actual	**Positivo**	TP (True Positive)	FN (False Negative)
	Negativo	FP (False Positive)	TN (True Negative)

Ilustración 4 Matriz de confusión, adaptado por (Melo 2024)

Además, existen otras métricas a la hora de evaluar un modelo de machine learning, accuracy, precision, recall, specificity y F1-score también evalúan los conceptos de TP, FP. FN.

- **Accuracy:** la métrica base utilizada para evaluar el modelo suele ser la Precisión, la cual analiza la cantidad de predicciones correctas entre todas las predicciones.
- **Precision:** mide cuántas de las predicciones positivas son correctas, es decir corresponde a los verdaderos positivos.
- **Recall:** recall o sensibilidad, mide cuántos casos de los casos positivos se predijeron correctamente, con especial énfasis en los casos positivos.
- **Specificity:** la especificidad es una medida de cuántas predicciones negativas hechas son correctas.
- **F1-score:** combina precision y recall y se la describe como la media armónica de ambas métricas. Proporciona una métrica única que combina precisión y recuperación, equilibrando el equilibrio entre ambas.

3.3. Algoritmos de Machine Learning

Tras la revisión de la literatura, se utilizará técnicas de aprendizaje supervisado de clasificación mediante algoritmos de regresión logística y algoritmos de clasificación mediante árboles de

decisión, al ser los más efectivos para predecir variables en el campo educativo, específicamente en la predicción de la deserción escolar.

3.4. Herramientas

Las herramientas para implementar estos algoritmos de regresión logística y árboles de decisión incluyen bibliotecas de programación y software estadístico como Python, que usa bibliotecas como Scikit-learn para entrenar los modelos de aprendizaje.

3.5. Diagrama del Modelo Predictivo

La ilustración 4 muestra el diagrama del modelo aplicado para esta investigación:

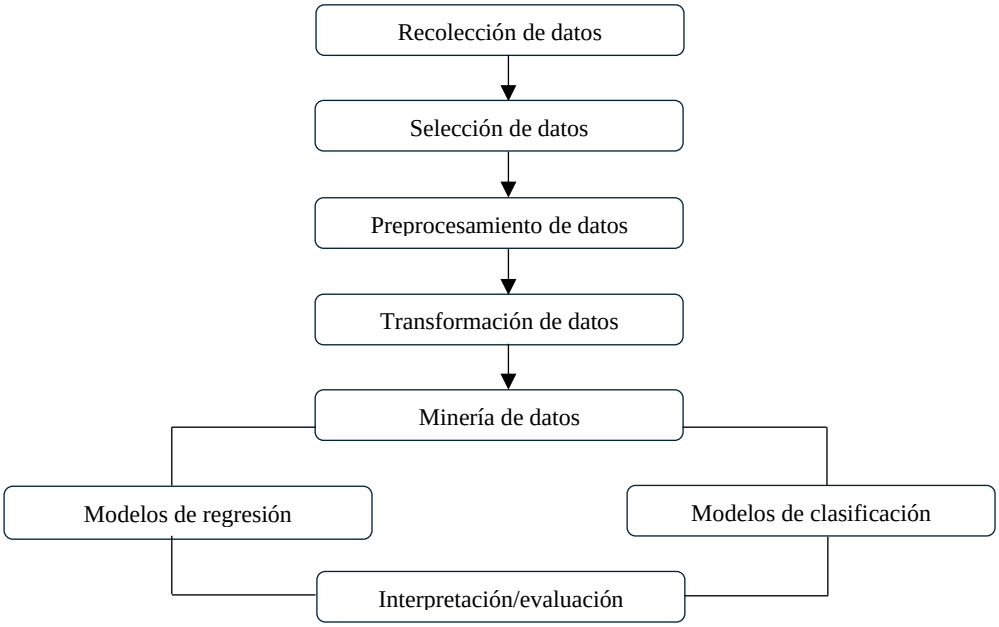


Ilustración 5 Diagrama del modelo predictivo (Timarán-Pereira, Hernández-Arteaga, Caicedo-Zambrano, Hidalgo-Troya, & Alvarado-Pérez, 2016), adaptado por (Melo 2024)

4. CAPÍTULO IV

El modelo predictivo de los factores de deserción escolar propuesto fue implementado utilizando la metodología KDD siguiendo todas las etapas de selección, preprocesamiento, transformación, minería de datos y evaluación e interpretación. El modelo se elaboró sobre un data set con características de las instituciones educativas a nivel nacional, tales como código de la provincia y cantón, código de la institución educativa, tipo de escolarización, tipo de educación, niveles educativos, clase de sostenimiento de la escuela, área donde está ubicada, número de docentes y administrativos por curso, modalidad educativa, tipo de jornada escolar, número de estudiantes matriculados, promovidos de año escolar, no promovidos de año escolar, desertores o que abandonaron sus estudios clasificados por género. El dataset fue obtenido del portal oficial del MINEDUC Datos Abiertos, de los registros administrativos del fin de cada

periodo escolar considerado para el análisis, es decir desde 2012-2013 al 2020-2021. Se limpió el dataset minuciosamente considerando los valores nulos, valores atípicos (outliers), y luego sometido a un análisis con Visual Studio Code y Python, aplicando los algoritmos de clasificación mediante regresión logística, árboles de decisión, random forest, K- Nearest Neighbors y Support Vector Machine.

4.1. Aplicación de la metodología KDD

El objetivo de este proyecto de investigación es determinar los factores de deserción escolar de los estudiantes de bachillerato del Ecuador. Para ello se propone usar la analítica de datos mediante el uso de algoritmos de aprendizaje supervisado, utilizando modelos de regresión logística, y se propone adicionalmente técnicas de clasificación de árboles de decisión, para poder comparar los resultados de estos algoritmos de aprendizaje supervisado.
La ilustración 5 muestra el proceso de KDD a seguir con el proyecto:

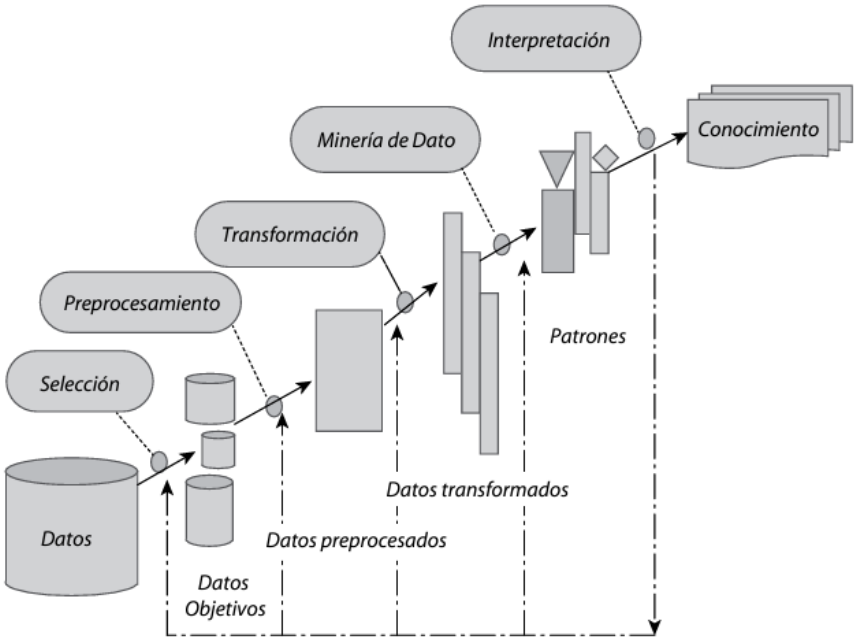


Ilustración 6 Etapas del proceso de KDD

Fuente: The Process of Knowledge Discovery on Databases

4.2. Etapa de selección

Para el proyecto de investigación, los datos se obtuvieron de los registros administrativos del portal web de datos abiertos del Ministerio de Educación. Los datos comprenden los periodos académicos del 2012 al 2021 considerando los registros al finalizar el año escolar[4] en formato .csv y la fuente de estos datos se obtiene del Archivo Maestro de Instituciones Educativas – AMIE.

[4] Información obtenida del portal de datos abiertos del Ministerio de Educación de Ecuador. La información a utilizar de fin de periodo escolar es una categorización asignada a los registros de estudiantes ingresados al término de cada periodo escolar.

La Tabla 1 muestra los periodos escolares de interés.

N°	Periodo escolar
1	Periodo 2012-2013 Fin
2	Periodo 2013-2014 Fin
3	Periodo 2014-2015 Fin
4	Periodo 2015-2016 Fin
5	Periodo 2016-2017 Fin
6	Periodo 2017-2018 Fin
7	Periodo 2018-2019 Fin
8	Periodo 2019-2020 Fin
9	Periodo 2020-2021 Fin

Tabla 1 Periodos escolares, adaptado por (Melo 2024)

4.2.1. Descripción de los datos

La muestra está conformada por atributos relevantes por institución educativa del Ecuador del periodo 2012 – 2021 a nivel nacional. Los datos fueron obtenidos de la página oficial del Ministerio de Educación del Ecuador (MINEDUC) donde se obtuvo 10 bases de datos para estos años. Posteriormente, se unificó las bases de datos en un solo dataset, para observar los datos en conjunto. Tras revisar toda la documentación sobre las cifras de deserción escolar en el país, conociendo que la deserción se concentra en el nivel de bachillerato, la investigación se centrará en datos de este nivel educativo. Se entiende a los estudiantes que cursan el nivel educativo de secundaria o bachillerato a adolescentes de 15 a 17 años de edad en los grados de 1ro a 3er de bachillerato a nivel nacional.

La Tabla 2 muestra las columnas del dataset concatenado.

CAMPO
Periodo
Zona
Cod_Provincia
Cod_Canton
Codigo_Institucion
Escolarizacion
Tipo_Educacion
Nivel_Educacion
Sostenimiento
Area
Regimen_Escolar
Jurisdiccion
Modalidad

Jornada
Tenencia_Inmueble_Edificio
Acceso_Edificio
Docentes_Femenino
Docentes_Masculino
Total_Docentes
Administrativos_Femenino
Administrativos_Masculino
Total_Administrativos
EstudiantesFemeninoPrimerAñoBACH
EstudiantesMasculinoPrimerAñoBACH
EstudiantesFemeninoPromovidosPrimerAñoBACH
EstudiantesMasculinoPromovidosPrimerAñoBACH
EstudiantesFemeninoNoPromovidosPrimerAñoBACH
EstudiantesMasculinoNoPromovidosPrimerAñoBACH
EstudiantesFemeninoDesertoresPrimerAñoBACH
EstudiantesMasculinoDesertoresPrimerAñoBACH
EstudiantesFemeninoNoActualizadosPrimerAñoBACH
EstudiantesMasculinoNoActualizadosPrimerAñoBACH
EstudiantesFemeninoSegundoAñoBACH
EstudiantesMasculinoSegundoAñoBACH
EstudiantesFemeninoPromovidosSegundoAñoBACH
EstudiantesMasculinoPromovidosSegundoAñoBACH
EstudiantesFemeninoNoPromovidosSegundoAñoBACH
EstudiantesMasculinoNoPromovidosSegundoAñoBACH
EstudiantesFemeninoDesertoresSegundoAñoBACH
EstudiantesMasculinoDesertoresSegundoAñoBACH
EstudiantesFemeninoNoActualizadosSegundoAñoBACH
EstudiantesMasculinoNoActualizadosSegundoAñoBACH
EstudiantesFemeninoTercerAñoBACH
EstudiantesMasculinoTercerAñoBACH
EstudiantesFemeninoPromovidosTercerAñoBACH
EstudiantesMasculinoPromovidosTercerAñoBACH
EstudiantesFemeninoNoPromovidosTercerAñoBACH
EstudiantesMasculinoNoPromovidosTercerAñoBACH
EstudiantesFemeninoDesertoresTercerAñoBACH
EstudiantesMasculinoDesertoresTercerAñoBACH
EstudiantesFemeninoNoActualizadoTercerAñoBACH
EstudiantesMasculinoNoActualizadoTercerAñoBACH

Tabla 2 Campos del dataset concatenado, adaptado por (Melo 2024)

El dataset presentado es el resultado de unir los 10 archivos .csv. Continuaremos en la siguiente etapa del proceso de KDD para el preprocesamiento y la limpieza del dataset. Previo al preprocesamiento y al revisar nuevamente el dataset se notó un error en la concatenación de los archivos csv ya que varios registros no se copiaron correctamente en las columnas, por ello se optó por arreglar el archivo y subirlo directamente en formato .xlsx.
En la Ilustración 7 se muestra el dataset concatenado con la información de los 10 archivos csv, como se observa en lenguaje de programación Python. El dataset presenta 52 columnas y 173.774 registros.

	Periodo	Zona	Cod_Provincia	Cod_Canton	Codigo_Institucion	Escolarizacion	Tipo_Educacion	Nivel_Educacion	Sostenimiento	Area	...
0	2012-2013 Fin	ZONA 6	1	101	01B00014	Escolarizada	EducaciÃ³n Regular	EducaciÃ³n BÃ¡sica	Particular	UrbanaINEC	...
1	2012-2013 Fin	ZONA 6	1	101	01B00020	Escolarizada	EducaciÃ³n Regular	EGB y Bachillerato	Fiscal	RuralINEC	...
2	2012-2013 Fin	ZONA 6	1	101	01B00021	Escolarizada	EducaciÃ³n Regular	EducaciÃ³n BÃ¡sica	Fiscal	RuralINEC	...
3	2012-2013 Fin	ZONA 6	1	101	01B00022	Escolarizada	EducaciÃ³n Regular	EGB y Bachillerato	Fiscal	UrbanaINEC	...
4	2012-2013 Fin	ZONA 6	1	101	01B00027	Escolarizada	EducaciÃ³n Regular	Inicial y EGB	Fiscal	RuralINEC	...

5 rows × 52 columns

Ilustración 7 Exploración del dataset, adaptado por (Melo 2024)

4.2.2. Etapa de preprocesamiento y/o limpieza

En esta etapa se inició explorando el formato de las variables para su posterior procesamiento. Se prosiguió con la revisión de la escritura de las etiquetas de cada variable, ya que en la exploración del dataset las etiquetas estaban mal escritas o el programa los leía como duplicados en la mayoría de los casos. También se actualizó el nombre de cada variable por uno más claro.

4.2.3. Identificación de variables numéricas y categóricas

Se identificaron 14 variables categóricas en el dataset: Periodo, Zona, Codigo_Institucion, Escolarizacion, Tipo_Educacion, Nivel_Educacion, Sostenimiento, Area, Regimen_Escolar, Jurisdiccion, Modalidad, Jornada, Tenencia_Inmueble_Edificio, Acceso_Edificio. Posteriormente se identificó las etiquetas de cada variable y la distribución de frecuencias de los valores de las variables categóricas. El número de etiquetas de una variable categórica se denomina cardinalidad. Un número elevado de etiquetas dentro de una variable se conoce como cardinalidad alta, este fenómeno puede plantear algunos problemas graves en el modelo.
La tabla 3 muestra la cardinalidad de las variables categóricas.

Variable categórica	Etiquetas
Periodo	contiene 9 etiquetas
Zona	contiene 10 etiquetas
Codigo_Institucion	contiene 28073 etiquetas
Escolarizacion	contiene 1 etiquetas
Tipo_Educacion	contiene 5 etiquetas
Nivel_Educacion	contiene 29 etiquetas
Sostenimiento	contiene 4 etiquetas
Area	contiene 2 etiquetas
Regimen_Escolar	contiene 3 etiquetas
Jurisdiccion	contiene 5 etiquetas
Modalidad	contiene 28 etiquetas
Jornada	contiene 7 etiquetas
Tenencia_Inmueble_Edificio	contiene 9 etiquetas
Acceso_Edificio	contiene 3 etiquetas

Tabla 3 Cardinalidad de variables categóricas, adaptado por (Melo 2024)

Al analizar la tabla se observa que las variables tienen una cardinalidad relativamente baja, a excepción de Codigo_Institucion, Nivel_Educación y Modalidad. Estas variables tienen alta cardinalidad pues el código de institución es el identificador único de las escuelas a nivel nacional.

Posteriormente se comprobó el número de variables numéricas dentro del dataset existen 38 variables numéricas. La tabla 4 muestra las variables numéricas.

Variable numérica
Cod_Provincia
Cod_Canton
Docentes_Femenino
Docentes_Masculino
Total_Docentes
Administrativos_Femenino
Administrativos_Masculino
Total_Administrativos
EstudiantesFemeninoPrimerAÃ±oBACH
EstudiantesMasculinoPrimerAÃ±oBACH
EstudiantesFemeninoPromovidosPrimerAÃ±oBACH
EstudiantesMasculinoPromovidosPrimerAÃ±oBACH
EstudiantesFemeninoNoPromovidosPrimerAÃ±oBACH
EstudiantesMasculinoNoPromovidosPrimerAÃ±oBACH
EstudiantesFemeninoDesertoresPrimerAÃ±oBACH

EstudiantesMasculinoDesertoresPrimerAÃ±oBACH
EstudiantesFemeninoNoActualizadosPrimerAÃ±oBACH
EstudiantesMasculinoNoActualizadosPrimerAÃ±oBACH
EstudiantesFemeninoSegundoAÃ±oBACH
EstudiantesMasculinoSegundoAÃ±oBACH
EstudiantesFemeninoPromovidosSegundoAÃ±oBACH
EstudiantesMasculinoPromovidosSegundoAÃ±oBACH
EstudiantesFemeninoNoPromovidosSegundoAÃ±oBACH
EstudiantesMasculinoNoPromovidosSegundoAÃ±oBACH
EstudiantesFemeninoDesertoresSegundoAÃ±oBACH
EstudiantesMasculinoDesertoresSegundoAÃ±oBACH
EstudiantesFemeninoNoActualizadosSegundoAÃ±oBACH
EstudiantesMasculinoNoActualizadosSegundoAÃ±oBACH
EstudiantesFemeninoTercerAÃ±oBACH
EstudiantesMasculinoTercerAÃ±oBACH
EstudiantesFemeninoPromovidosTercerAÃ±oBACH
EstudiantesMasculinoPromovidosTercerAÃ±oBACH
EstudiantesFemeninoNoPromovidosTercerAÃ±oBACH
EstudiantesMasculinoNoPromovidosTercerAÃ±oBACH
EstudiantesFemeninoDesertoresTercerAÃ±oBACH
EstudiantesMasculinoDesertoresTercerAÃ±oBACH
EstudiantesFemeninoNoActualizadoTercerAÃ±oBACH
EstudiantesMasculinoNoActualizadoTercerAÃ±oBACH

Tabla 4 Variables numéricas, adaptado por (Melo 2024)

Después de identificar las variables numéricas y categóricas, se revisó los nombres de las etiquetas de cada uno y se actualizaron a nombres más cortos y entendibles. Se actualizaron las etiquetas de Periodo, Zona, Cod_Provincia, tipo_education, Nivel_Educaion, Jurisdicción, Modalidad, Jornada, Tenencia_Inmueble_Edificio, Acceso_Edificio, Area.
La ilustración 8 muestra reflejados estos cambios en las etiquetas de las variables.

	Periodo	Zona	Cod_Provincia	Cod_Canton	Codigo_Institucion	Escolarizacion	Tipo_Educacion	Nivel_Educacion	Sostenimiento	Area	...
0	2012-2013	6	1	101	01B00014	Escolarizada	Educacion Regular	Educacion Basica	Particular	Urbana	...
1	2012-2013	6	1	101	01B00020	Escolarizada	Educacion Regular	EGB y Bachillerato	Fiscal	Rural	...
2	2012-2013	6	1	101	01B00021	Escolarizada	Educacion Regular	Educacion Basica	Fiscal	Rural	...
3	2012-2013	6	1	101	01B00022	Escolarizada	Educacion Regular	EGB y Bachillerato	Fiscal	Urbana	...
4	2012-2013	6	1	101	01B00027	Escolarizada	Educacion Regular	Inicial y EGB	Fiscal	Rural	...

5 rows × 52 columns

Ilustración 8 Dataset con cambio en nombre de etiquetas, adaptado por (Melo 2024)

Luego se comenzó el proceso de actualizar los nombres de las variables a un nombre más entendible. Se modificó la mayor parte de las variables ya que los nombres venían con errores de tildes o eran demasiado largos.
La ilustración 9 muestra el cambio de nombre de las variables.

	Periodo	Zona	Cod_Provincia	Cod_Canton	Codigo_Institucion	Escolarizacion	Tipo_Educacion	Nivel_Educacion	Sostenimiento	Area	...	TerceroF
0	2012-2013	6	1	101	01B00014	Escolarizada	Educacion Regular	Educacion Basica	Particular	Urbana	...	0.0
1	2012-2013	6	1	101	01B00020	Escolarizada	Educacion Regular	EGB y Bachillerato	Fiscal	Rural	...	13.0
2	2012-2013	6	1	101	01B00021	Escolarizada	Educacion Regular	Educacion Basica	Fiscal	Rural	...	0.0

3 rows × 52 columns

Ilustración 9 Dataset con cambio en nombre de variables, adaptado por (Melo 2024)

La Tabla 5 muestra los nombres de variables, su descripción, el tipo de variable y el rango.

Variable	Descripción	Tipo	Rango
Periodo	Periodo escolar codificado para registros inicio o fin. Fin del periodo escolar se refiere a la información estadística al culminar el periodo escolar, el cual determina la descomposición de la matrícula (Promovidos, No promovidos, Abandono).	Discreta	2012-2013 2013-2014 2014-2015 2015-2016 2016-2017 2017-2018 2018-2019 2019-2020 2020-2021
Zona	Zona de planificación, según ubicación geográfica de la institución educativa	Nominal	1, 2, 3, 4, 5, 6, 7, 8, 9, NO DELIMITADA
Cod_Provincia	Determina el código de la provincia de donde proviene el estudiante	Nominal	01 02 03 .... 024 090
Cod_Canton	Código del cantón en la que se encuentra ubicada la institución educativa	Nominal	0101 0102 ... 9004
Tipo_Educacion	Servicio educativo que oferta la institución.	Nominal	Educación Regular, Educación Especial, Formación Artística

Nivel_educacion	Determina el nivel educativo del estudiante	Nominal	Inicial, Educación Básica, Bachillerato, Alfabetización P.P., Artesanal P.P.
Sostenimiento	Clasifica a la institución educativa, conforme la proveniencia de los recursos que obtiene	Nominal	Fiscal Fiscomisional Municipal Particular
Area	Corresponde al área geográfica donde se encuentra estudiando	Nominal	Urbana Rural
Regimen_escolar	Determina el régimen escolar que siguió el estudiante	Nominal	Costa, Sierra, Permanente
Jurisdiccion	Identifica a las instituciones educativas conforme su administración	Nominal	Intercultural Intercultural Bilingüe
Modalidad	Establece el tipo de dinámica de trabajo mediante el cual se realiza el proceso de enseñanza y aprendizaje, determinado por las estrategias y formas de comunicación empleadas. La institución puede ofertar una o varias modalidades.	Nominal	Presencial Semipresencial A distancia Educación abierta ...
Jornada	Establece la proporción del día en que los estudiantes y docentes interactúan en el proceso de enseñanza aprendizaje (la institución educativa puede ofertar una o más de una jornada en el mismo periodo (Matutina, Vespertina, Nocturna)	Nominal	Matutina Vespertina Nocturna
TenenciaI	Si la institución educativa es propia, comodato, arriendo, no conoce, cesión de derechos, invasión, prestado o propio.	Nominal	Arriendo Cesión de derechos Comodato Invasión No conoce Prestado

			Propio
AccesoE		Nominal	Aéreo Fluvial Terrestre
DocentesF	Número de docentes por sexo femenino, registrados por la institución educativa	Discreta	
DocentesM	Número de docentes por sexo masculino, registrados por la institución educativa	Discreta	
DocentesT	Número total de docentes del Sistema Educativo Nacional, registrados por la institución educativa	Discreta	
AdminF	Número de administrativos por sexo femenino, registrados por la institución educativa	Discreta	
AdminM	Número de administrativos por sexo masculino, registrados por la institución educativa	Discreta	
AdminT	Número total de administrativos del Sistema Educativo Nacional, registrados por la institución educativa	Discreta	
PimeroF	Número de estudiantes mujeres registradas en primer año de Bachillerato	Discreta	
PrimeroM	Número de estudiantes de 1ero a 3ro de Bachillerato, registrados por la institución educativa	Discreta	
PrimeroF_Prom	Número de estudiantes mujeres que finalizaron el periodo escolar, cumpliendo con todos los requisitos para ser promovidos al grado siguiente del nivel educativo correspondiente	Discreta	
PrimeroM_Prom	Número de estudiantes hombres que finalizaron el periodo escolar, cumpliendo con todos los	Discreta	

	requisitos para ser promovidos al grado siguiente del nivel educativo correspondiente		
PrimeroF_NoProm	Número de estudiantes mujeres que finalizaron el periodo escolar y que no cumplieron con los requisitos para ser promovidos al grado siguiente del nivel educativo correspondiente	Discreta	
PrimeroM_NoProm	Número de estudiantes hombres que finalizaron el periodo escolar y que no cumplieron con los requisitos para ser promovidos al grado siguiente del nivel educativo correspondiente	Discreta	
PrimeroF_Deser	Número de estudiantes mujeres contabilizados al final de un periodo escolar que abandonan el primer año de bachillerato	Discreta	
PrimeroM_Deser	Número de estudiantes hombres contabilizados al final de un periodo escolar que abandonan el primer año de bachillerato	Discreta	
SegundoF	Número de estudiantes mujeres registradas en segundo año de Bachillerato	Discreta	
SegundoM	Número de estudiantes mujeres registradas en segundo año de Bachillerato	Discreta	
SegundoF_Prom	Número de estudiantes mujeres que finalizaron el periodo escolar, cumpliendo con todos los requisitos para ser promovidos al grado siguiente del nivel educativo correspondiente	Discreta	
SegundoM_Prom	Número de estudiantes hombres que finalizaron el periodo escolar, cumpliendo con todos los requisitos para ser promovidos al	Discreta	

	grado siguiente del nivel educativo correspondiente		
SegundoF_NoProm	Número de estudiantes mujeres que finalizaron el periodo escolar y que no cumplieron con los requisitos para ser promovidos al grado siguiente del nivel educativo correspondiente	Discreta	
SegundoM_NoProm	Número de estudiantes hombres que finalizaron el periodo escolar y que no cumplieron con los requisitos para ser promovidos al grado siguiente del nivel educativo correspondiente	Discreta	
SegundoF_Deser	Número de estudiantes mujeres contabilizados al final de un periodo escolar que abandonan el segundo año de bachillerato	Discreta	
SegundoM_Deser	Número de estudiantes hombres contabilizados al final de un periodo escolar que abandonan el segundo año de bachillerato	Discreta	
TerceroF	Número de estudiantes mujeres registradas en tercer año de Bachillerato	Discreta	
TerceroM	Número de estudiantes hombres registradas en tercer año de Bachillerato	Discreta	
TerceroF_Prom	Número de estudiantes mujeres que finalizaron el periodo escolar, cumpliendo con todos los requisitos para ser promovidos al grado siguiente del nivel educativo correspondiente	Discreta	
TerceroM_Prom	Número de estudiantes hombres que finalizaron el periodo escolar, cumpliendo con todos los requisitos para ser promovidos al grado siguiente del nivel educativo correspondiente	Discreta	

TerceroF_NoProm	Número de estudiantes mujeres que finalizaron el periodo escolar y que no cumplieron con los requisitos para ser promovidos al grado siguiente del nivel educativo correspondiente	Discreta	
TerceroM_NoProm	Número de estudiantes hombres que finalizaron el periodo escolar y que no cumplieron con los requisitos para ser promovidos al grado siguiente del nivel educativo correspondiente	Discreta	
TerceroF_Deser	Número de estudiantes mujeres contabilizados al final de un periodo escolar que abandonan el tercer año de bachillerato	Discreta	
TerceroM_Deser	Número de estudiantes hombres contabilizados al final de un periodo escolar que abandonan el tercer año de bachillerato	Discreta	

Ilustración 10 Variables, descripción y tipo de variable, adaptado por (Melo 2024)

4.2.4. Verificación de valores perdidos o NaN

Se identificaron 37 valores perdidos correspondientes a las variables de estudiantes de bachillerato de primero, segundo y tercer año. Se decide eliminar estos registros pues al parecer son datos que no fueron capturados al momento de realizar el levantamiento de la información por las autoridades competentes, por tanto no se tiene registro de esta información. Además, considerando que no es un número significativo para el dataset se entiende que no afectará los resultados de análisis posteriores.

La Ilustración 11 muestra el dataset con estas modificaciones. Se observa que ahora se cuenta con 52 columnas y 173.737 registros.

```
df_final.head()
```
✓ 0.0s 📊 Open 'df_final' in Data Wrangler

	Periodo	Zona	Cod_Provincia	Cod_Canton	Codigo_Institucion	Escolarizacion	Tipo_Educacion	Nivel_Educacion	Sostenimiento	Area	...	TerceroF	TerceroM
0	2012-2013	6	1	101	01B00014	Escolarizada	Educacion Regular	Educacion Basica	Particular	Urbana	...	0.0	0.0
1	2012-2013	6	1	101	01B00020	Escolarizada	Educacion Regular	EGB y Bachillerato	Fiscal	Rural	...	13.0	12.0
2	2012-2013	6	1	101	01B00021	Escolarizada	Educacion Regular	Educacion Basica	Fiscal	Rural	...	0.0	0.0
3	2012-2013	6	1	101	01B00022	Escolarizada	Educacion Regular	EGB y Bachillerato	Fiscal	Urbana	...	36.0	33.0
4	2012-2013	6	1	101	01B00027	Escolarizada	Educacion Regular	Inicial y EGB	Fiscal	Rural	...	0.0	0.0

5 rows × 52 columns

Ilustración 11 Exploración del dataset modificado, adaptado por (Melo 2024)

La ilustración 12 muestra la información general del dataset el cual contiene 52 variables de tipo float63, int64 y 11 tipo objeto o categórica.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 173774 entries, 0 to 173773
Columns: 52 entries, Periodo to EstudiantesMasculinoNoActualizadoTercerAÃ±oBACH
dtypes: float64(30), int64(9), object(13)
memory usage: 68.9+ MB
```

Ilustración 12 Información del dataset, adaptado por (Melo 2024)

4.2.5. Formateo de tipos de datos

Se realizó un análisis del tipo de variable para recodificar al formato adecuado. Tras el análisis se identificó que las variables numéricas que comprenden el número de estudiantes por año de bachillerato (primer año, segundo año y tercer año) registrados, promovidos, no promovidos y desertores eran de tipo variable numérica continua (float) por lo que se modificó a tipo de variable discretas (int64).

La ilustración 13 muestra el tipo de variable numérica float, int64 en el dataset.

PrimeroF	int64
PrimeroM	int64
PrimeroF_Prom	int64
PrimeroM_Prom	int64
PrimeroF_NoProm	int64
PrimeroM_NoProm	int64
PrimeroF_Deser	int64
PrimeroM_Deser	int64
EstudiantesFemeninoNoActualizadosPrimerAÃ±oBACH	float64
EstudiantesMasculinoNoActualizadosPrimerAÃ±oBACH	float64
SegundoF	int64
SegundoM	int64
SegundoF_Prom	int64
SegundoM_Prom	int64
SegundoF_NoProm	int64
SegundoM_NoProm	int64
SegundoF_Deser	int64
SegundoM_Deser	int64
EstudiantesFemeninoNoActualizadosSegundoAÃ±oBACH	float64
EstudiantesMasculinoNoActualizadosSegundoAÃ±oBACH	float64
TerceroF	int64
TerceroM	int64
TerceroF_Prom	int64
TerceroM_Prom	int64
TerceroF_NoProm	int64
TerceroM_NoProm	int64
TerceroF_Deser	int64
TerceroM_Deser	int64
EstudiantesFemeninoNoActualizadoTercerAÃ±oBACH	float64
EstudiantesMasculinoNoActualizadoTercerAÃ±oBACH	float64

Ilustración 13 Cambio de tipo de variable, adaptado por (Melo 2024)

La ilustración 14 muestra este cambio en las variables numéricas de estudiantes de bachillerato.

PrimeroF	int64
PrimeroM	int64
PrimeroF_Prom	int64
PrimeroM_Prom	int64
PrimeroF_NoProm	int64
PrimeroM_NoProm	int64
PrimeroF_Deser	int64
PrimeroM_Deser	int64
EstudiantesFemeninoNoActualizadosPrimerAÃ±oBACH	float64
EstudiantesMasculinoNoActualizadosPrimerAÃ±oBACH	float64
SegundoF	int64
SegundoM	int64
SegundoF_Prom	int64
SegundoM_Prom	int64
SegundoF_NoProm	int64
SegundoM_NoProm	int64
SegundoF_Deser	int64
SegundoM_Deser	int64
EstudiantesFemeninoNoActualizadosSegundoAÃ±oBACH	float64
EstudiantesMasculinoNoActualizadosSegundoAÃ±oBACH	float64
TerceroF	int64
TerceroM	int64
TerceroF_Prom	int64
TerceroM_Prom	int64
TerceroF_NoProm	int64
TerceroM_NoProm	int64
TerceroF_Deser	int64
TerceroM_Deser	int64
EstudiantesFemeninoNoActualizadoTercerAÃ±oBACH	float64
EstudiantesMasculinoNoActualizadoTercerAÃ±oBACH	float64

Ilustración 14 Variables numéricas modificadas, adaptado por (Melo 2024)

4.2.6. Revisión de datos ruidosos

Para revisar los datos con outliers o ruido en el dataset, primero seleccionamos aquellos que son datos numéricos. La ilustración 15 muestra las estadísticas descriptivas principales de este conjunto de datos numérico.

	Zona	Cod_Provincia	Cod_Canton	DocentesF	DocentesM	DocentesT	AdminF	AdminM	AdminT	PrimeroF
count	173737.000000	173737.000000	173737.000000	173737.000000	173737.000000	173737.000000	173737.000000	173737.000000	173737.000000	173737.000000
mean	4.985990	11.741276	1178.921807	7.910077	3.335219	11.245296	1.009163	0.675176	1.684339	8.256359
std	2.375596	7.210821	720.642710	13.066513	6.595340	18.599610	3.170490	2.290347	5.174159	32.109964
min	1.000000	1.000000	101.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	3.000000	8.000000	806.000000	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000
50%	5.000000	11.000000	1112.000000	3.000000	1.000000	4.000000	0.000000	0.000000	0.000000	0.000000
75%	7.000000	15.000000	1503.000000	9.000000	4.000000	13.000000	1.000000	0.000000	1.000000	0.000000
max	10.000000	90.000000	9004.000000	256.000000	164.000000	357.000000	107.000000	67.000000	170.000000	1107.000000

Ilustración 15 Estadísticos descriptivos del dataset, adaptado por (Melo 2024)

Los valores máximos son indicio de que pueden existir outliers en las variables DocentesF, DocentesM, AdminF, AdminM. Se analizará los outliers mediante gráficos boxplot para ver los cuantiles y los outliers.

La ilustración 16 muestra el gráfico boxplot de cada una de estas variables numéricas.

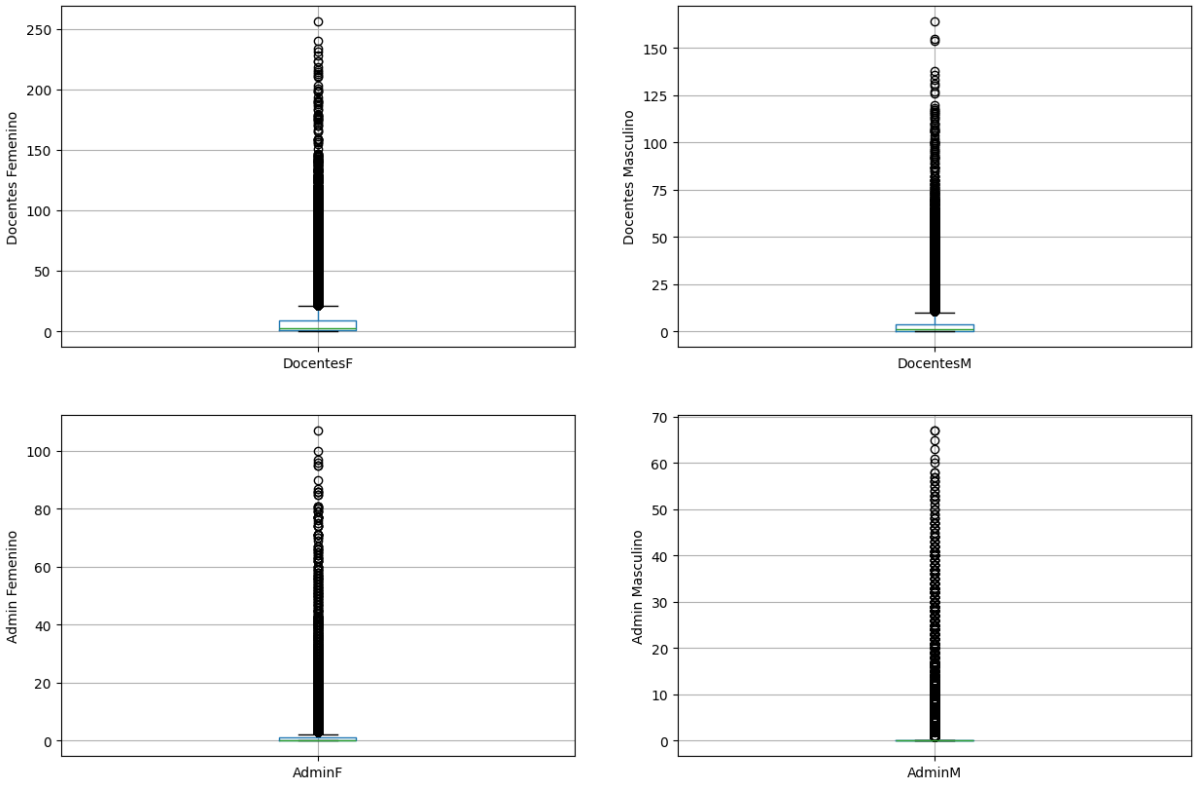


Ilustración 16 Boxplot para anális de outliers, adaptado por (Melo 2024)

Mediante el gráfico observamos que existe una alta cantidad de outliers dentro de estas variables. Por lo tanto, vamos a continuar analizando la distribución de estas variables para saber si existe sesgo.

La ilustración 17 muestra la distribución de las variables docentes y administrativos por género.

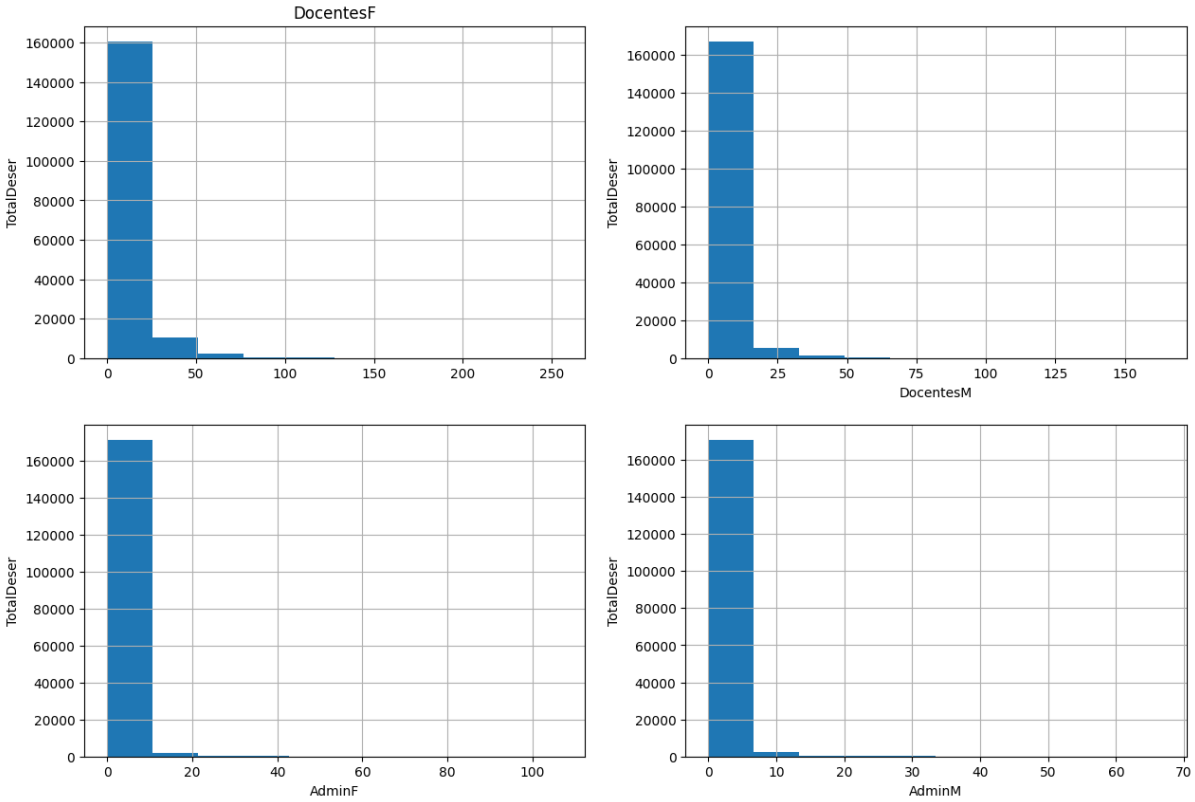


Ilustración 17 Distribución de variables, adaptado por (Melo 2024)

Mediante el gráfico se determina que tanto la distribución del número de docentes femenino, masculino y administrativos femenino y masculino tienen asimetría positiva (skewness).
Para determinar los outliers se utilizará el método z score, para determinar estos valores de manera efectiva.

Al analizar las variables DocentesF el límite superior encontrado fue de 47.10 por lo que los demás valores fuera de este límite se consideran como outliers. El dataset muestra que hay 3.698 filas con outliers. En este caso, la variable DocentesF representa docentes mujeres que trabajan en escuelas con más de 900 estudiantes, llegando a tener hasta 2.000 estudiantes según lo revisado en el dataset de outliers con esta variable, por lo que no se considera como un outlier tener más de 47 docentes mujeres ya que la escuela requiere de una mayor cantidad de docentes para dar clases a los estudiantes matriculados en estas escuelas.[5]
Respecto a la variable DocentesM, el límite superior encontrado es de 23.12, por lo que valores superiores a este límite se consideran como outliers. Se realizó un análisis similar a la variable DocentesF, y se identificaron 3.708 outliers en DocentesM. Al igual que el caso anterior, los

[5] El MINEDUC publicó un documento donde se registra a las instituciones educativas del país que tienen más de 900 estudiantes, con el código de la institución se pudo corroborar que los outliers identificados en la variable DocentesF pertenecen a esta clase de escuelas.

outliers que se muestran pertenecen a escuelas con un gran número de estudiantes y en su mayoría que ofertan más de un nivel educativo, por ello tienen mayor cantidad de docentes.

Para la variable AdminF se identificó un límite superior de 10.52, y se identificaron 2.665 outliers. Al analizar el sostenimiento en el que estos administrativos cumplían sus funciones se pudo corroborar que 1.919 de estos outliers pertenecen a escuelas con sostenimiento particular. Este análisis a nivel de sostenimiento con el dataset de outliers de la variable AdminF nos permite observar que la mayoría de las administradoras trabajan en escuelas particulares (72%).

4.3. Etapa de transformación o reducción

En la etapa de transformación se eliminaron variables que no aportaban a los modelos, como el código del cantón, el código de la institución (Cod_institucion) ya que no se tiene datos de todos los años de las escuelas, puede ser el caso de que una escuela cerró en un periodo escolar específico y ya no se tenga la información completa de esta escuela; además que el análisis de datos se hará en base a las características específicas de las escuelas. También se eliminó la variable escolarización ya que todas las instituciones educativas del dataset se encuentran escolarizadas como se vio en el análisis de la cardinalidad.

La educación escolarizada es acumulativa, progresiva, conlleva a la obtención de un título o certificado, tiene un año lectivo cuya duración se definirá técnicamente en el respectivo reglamento; responde a estándares y currículos específicos definidos por la Autoridad Educativa en concordancia con el Plan Nacional de Educación; y, brinda la oportunidad de formación y desarrollo de las y los ciudadanos dentro de los niveles inicial, básico y bachillerato (Ministerio de Educación de Ecuador, 2015).

Se eliminó las variables de estudiantes no actualizados en bachillerato, pues no se cuenta con información sobre qué representa esta variable en los registros administrativos del MINEDUC. Una vez eliminadas las variables se creó el dataset que contenga solo filas del nivel educativo bachillerato, lo que redujo las dimensiones del dataset.
La ilustración 18 muestra el dataset con estas modificaciones.

	Periodo	Zona	Cod_Provincia	Tipo_Educacion	Nivel_Educacion	Sostenimiento	Area	Regimen_Escolar	Jurisdiccion	Modalidad	...
1	2012-2013	6	1	Educacion Regular	EGB y Bachillerato	Fiscal	Rural	Sierra	Bilingue	Presencial y Semipresencial	...
3	2012-2013	6	1	Educacion Regular	EGB y Bachillerato	Fiscal	Urbana	Sierra	Bilingue	Semipresencial	...
26	2012-2013	6	1	Educacion Regular	EGB y Bachillerato	Fiscal	Urbana	Sierra	Bilingue	Presencial	...
29	2012-2013	6	1	Educacion Regular	EGB y Bachillerato	Fiscal	Urbana	Sierra	Bilingue	Presencial	...
43	2012-2013	6	1	Educacion Regular	Inicial, Educacion Basica y Bachillerato	Particular	Urbana	Sierra	Hispana	Presencial	...

5 rows × 43 columns

Ilustración 18 Dataset nivel bachillerato, adaptado por (Melo 2024)

El dataset contiene 43 variables, 32 de ellas tipo numérico (int64) y 11 de ellas tipo categóricas, consta de 35.810 registros.

4.3.1. Visualización de variables

La ilustración 19 muestra el tipo de sostenimiento de las instituciones educativas que ofrecen bachillerato en Ecuador. De acuerdo al gráfico de barras, vemos que la educación regular o también llamada ordinaria es la que predomina en el país, de acuerdo con el artículo 27 del Reglamento General a la LOEI, esta es la oferta para estudiantes que asisten regularmente a clases y están en la edad en correspondencia al grado cursado.

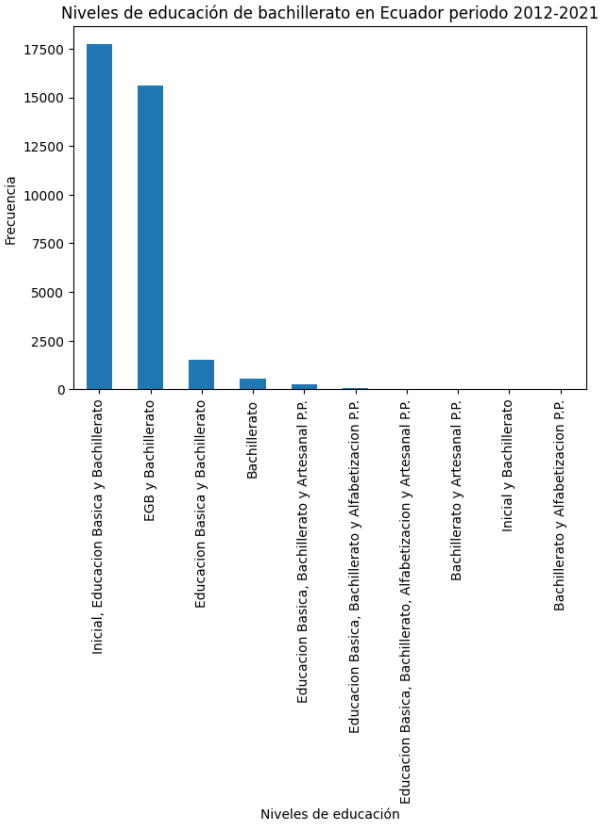


Ilustración 19 Niveles de educación de bachillerato en Ecuador, adaptado por (Melo 2024)

La ilustración 20 muestra el tipo de sostenimiento de las instituciones educativas en el nivel de bachillerato. Vemos en el gráfico que la mayor parte de instituciones educativas se encuentran financiadas por el Gobierno Central, le siguen las escuelas particulares que son construidas por personas naturales o jurídicas, las fiscomisionales que son instituciones de derecho privado con apoyo estatal y las instituciones educativas municipales, cuya fuente de financiamiento proviene de los GADs.

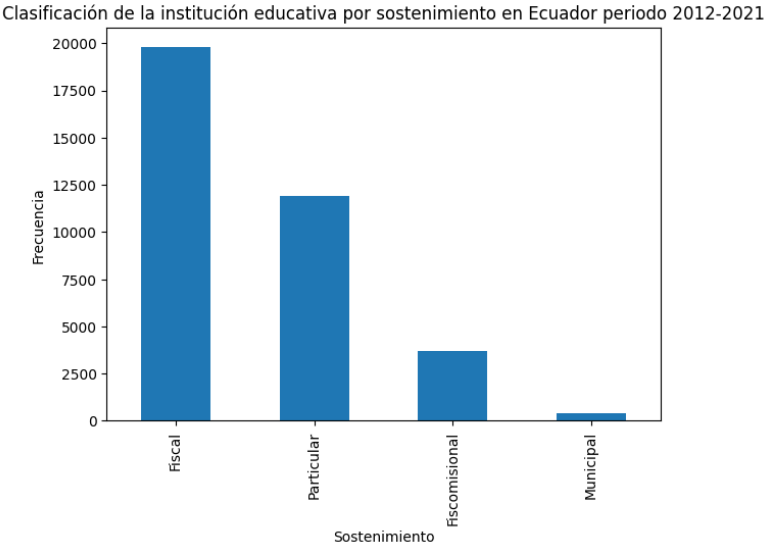


Ilustración 20 Sostenimiento de las instituciones educativas en Ecuador, adaptado por (Melo 2024)

La ilustración 21 muestra la distribución de la ubicación geográfica de las escuelas, donde se observa que la mayoría están situadas en el área urbana.

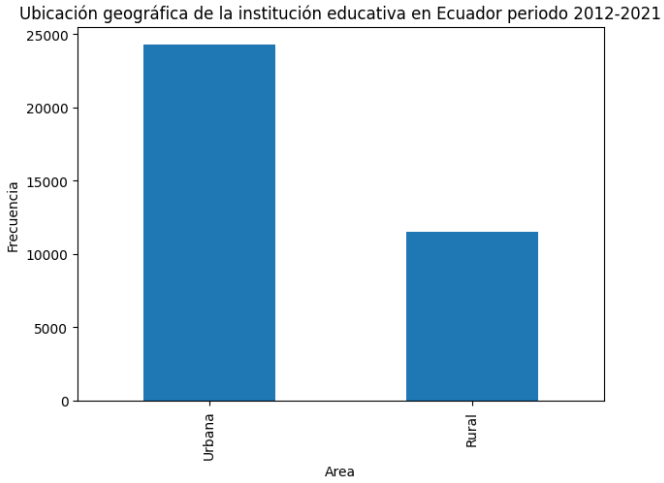


Ilustración 21 Ubicación geográfica de las instituciones educativas en Ecuador, adaptado por (Melo 2024)

La ilustración 22 muestra la distribución del régimen escolar de las instituciones educativas, donde su mayoría pertenecen al régimen costa.

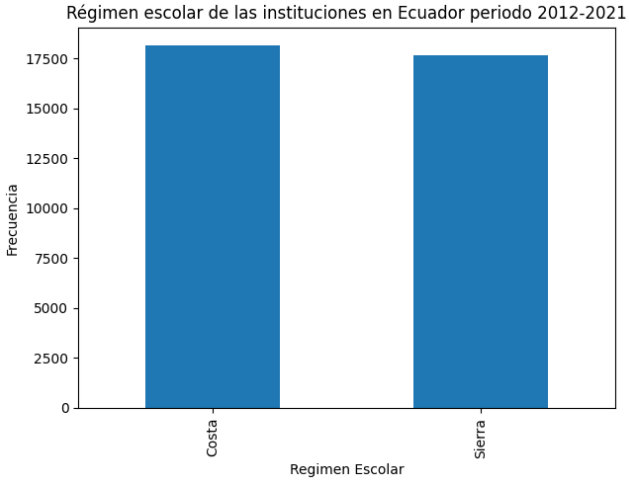


Ilustración 22 Régimen escolar de las instituciones educativas en Ecuador, adaptado por (Melo 2024)

La ilustración 23 muestra el tipo de jurisdicción de las instituciones educativas, donde predomina la administración hispana, seguido de jurisdicción intercultural.

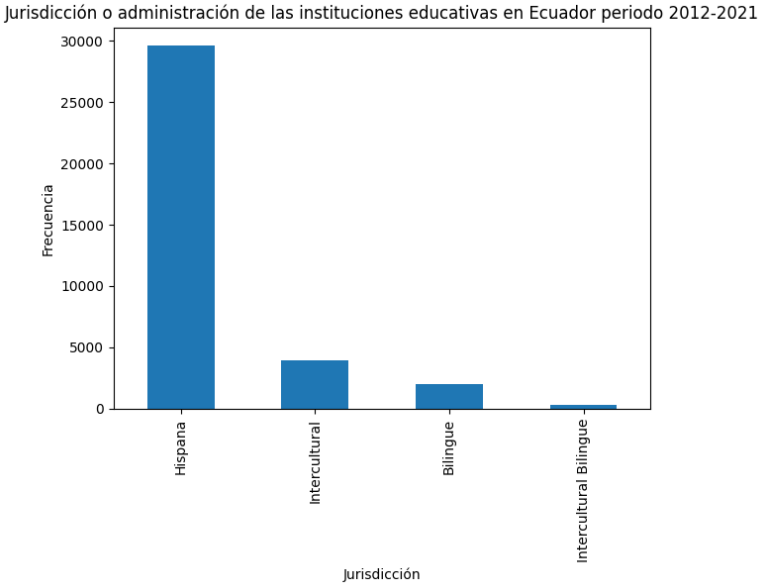


Ilustración 23 Tipo de jurisdicción de las instituciones educativas en Ecuador, adaptado por (Melo 2024)

La ilustración 24 muestra el tipo de modalidad educativa en las instituciones, en su mayoría es presencial en las instituciones educativas. Lo que se observa también es que existen diversas modalidades a distancia ofertadas formalmente por las instituciones, esto puede deberse a la respuesta educativa que brindó el MINEDUC en su momento ante la emergencia por Covid-19, para garantizar que los estudiantes continúen asistiendo a clases y ofertando los servicios educativos en diferentes modalidades.

Cabe destacar que en muchas localidades rurales no era posible mantener una modalidad a distancia por las desigualdades socioeconómicas que las familias y por ende, los estudiantes enfrentaban, por lo que se optó por ofrecer modalidades como educación en casa a través de material impreso, modalidad por radio o radiofónica y educación abierta.

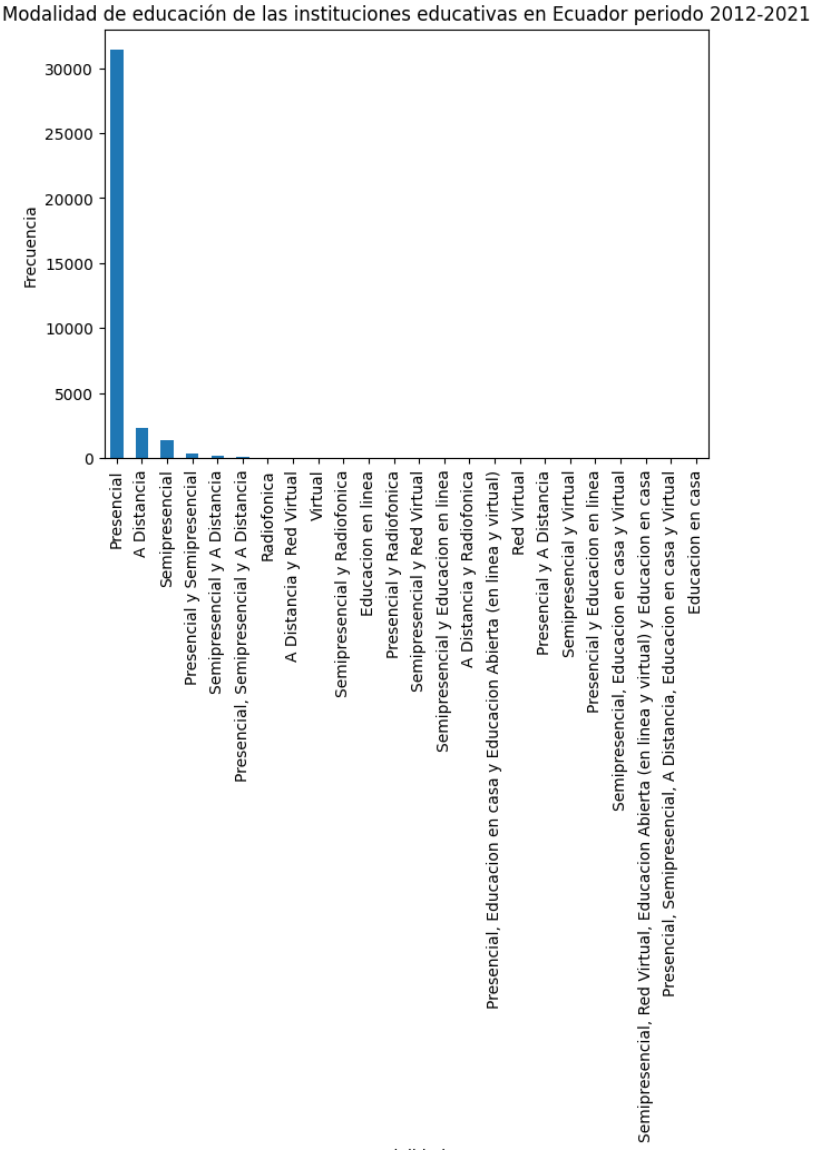


Ilustración 24 Modalidad de las instituciones educativas en Ecuador, adaptado por (Melo 2024)

La ilustración 20 muestra el tipo de tenencia del edificio, siendo en su mayoría de tenencia propia.

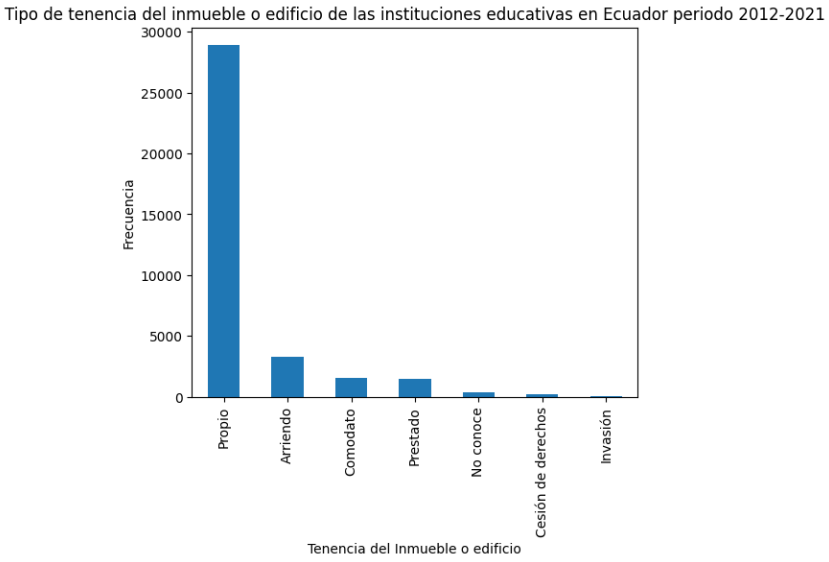


Ilustración 25 Tipo de tenencia de edificio escolar en Ecuador, adaptado por (Melo 2024)

La ilustración 26 muestra la forma de acceso a los edificios escolares, mayormente el acceso es por vía terrestre, sin embargo, hay otros edificios donde se debe acceder por agua o aire.

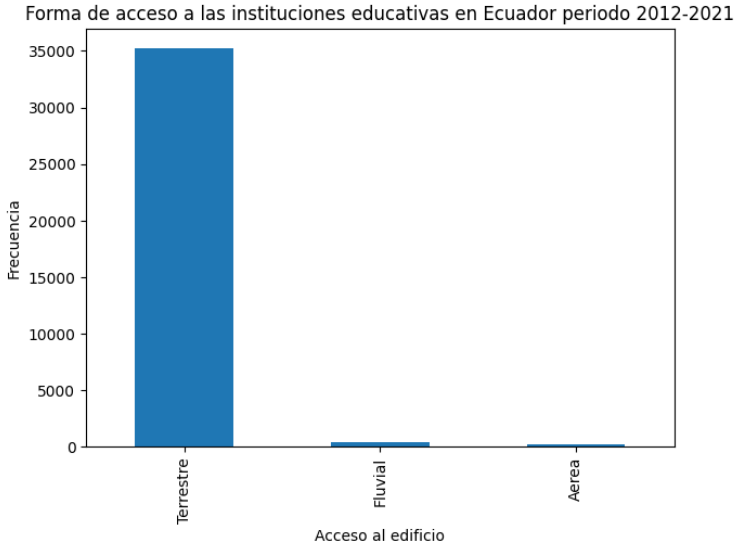


Ilustración 26 Forma de acceso al edificio escolar en Ecuador, adaptado por (Melo 2024)

En la ilustración 27 se observa el número de docentes en Ecuador por género, vemos que hay más docentes del género femenino que masculino. Sin embargo a partir del periodo escolar 2019-2020 hubo una disminución del número de docentes de género femenino.

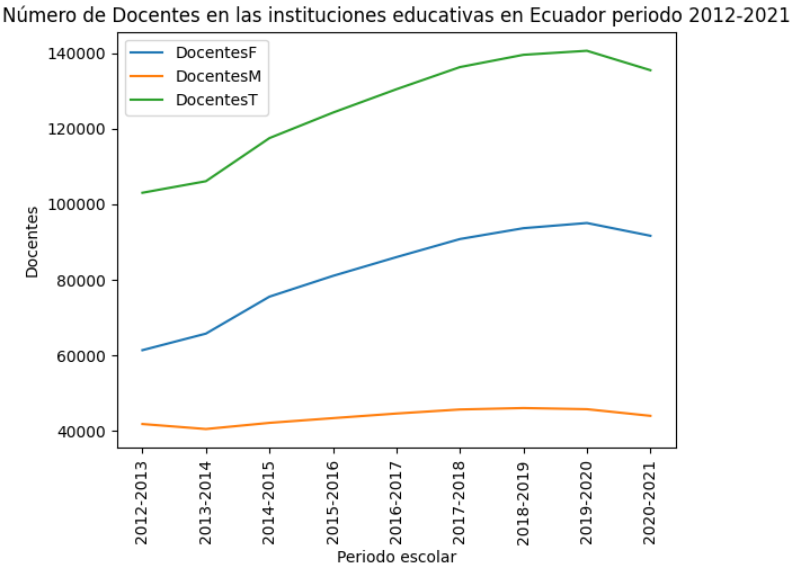


Ilustración 27 Docentes por género en Ecuador, adaptado por (Melo 2024)

Luego de revisar el dataset nuevamente, se decide eliminar la variable Nivel_Educacion pues los registros del dataset filtrado por nivel ya nos muestra el número de estudiantes y las escuelas de bachillerato. Posteriormente se creó nuevas variables utilizando el número de estudiantes matriculados, promovidos, no promovidos y desertores por año, por una variable que englobe todos los años de bachillerato.

La ilustración 28 muestra la creación de estas nuevas variables que se han incluido en el dataset.

TotalBachF	TotalBachM	TotalBach	TotalProm	TotalPromF	TotalPromM	TotalNoProm	TotalNoPromF	TotalNoPromM	TotalDeser
60	50	110	88	52	36	0	0	0	22
103	93	196	144	74	70	27	17	10	25
17	29	46	33	14	19	3	1	2	10
29	27	56	52	26	26	0	0	0	3
14	22	36	35	14	21	1	0	1	0

Ilustración 28 Nuevas variables estudiantes de bachillerato, adaptado por (Melo 2024)

La ilustración 29 muestra el número de estudiantes matriculados, promovidos, no promovidos en los diferentes periodos escolares. Se observa que la proporción de los promovidos es baja en los tres primeros periodos escolares analizados, sin embargo comienza a tener una tendencia creciente en el periodo 2015-2016. Mientras que los estudiantes no promovidos son pocos pero tienen una tendencia a la alta, excepto en el periodo 2019-2020, donde casi no existen estudiantes no promovidos, ya que en esta época por la pandemia, el MINEDUC decidió dar facilitades para que los estudiantes pasen al siguiente año escolar que les corresponde.

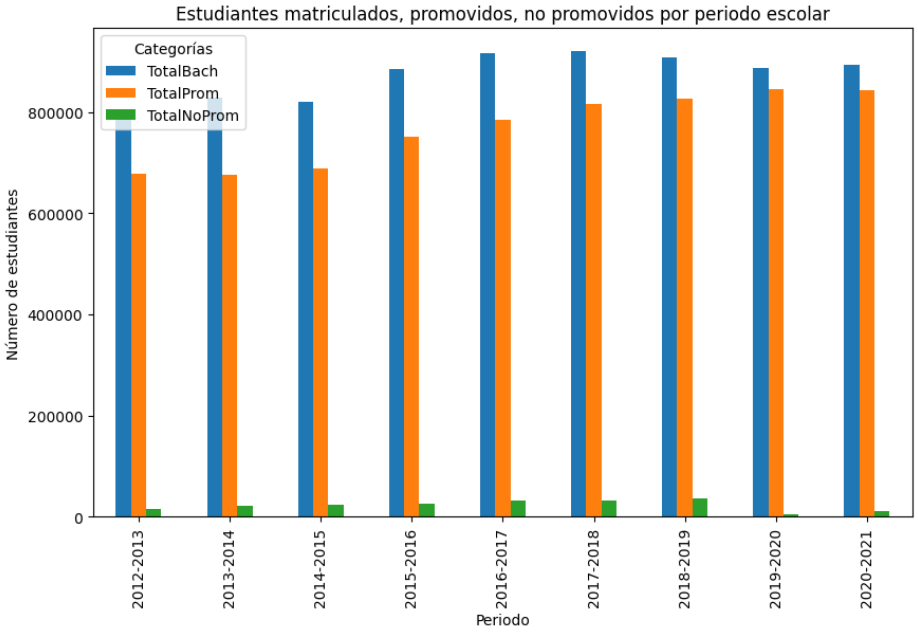


Ilustración 29 Distribución de estudiantes de bachillerato, adaptado por (Melo 2024)

La ilustración 30 muestra el número de estudiantes que desertaron en sus estudios en los periodos del estudio. En el gráfico podemos observar el número de estudiantes que han abandonado sus estudios entre los periodos escolares 2012-2013 al 2020-2021. Se observa que a partir del periodo escolar 2016-2017 existe una caída de la deserción. Sin embargo a partir del 2019-2020 podemos ver un leve incremento de la deserción, que puede deberse mayormente a la emergencia sanitaria por Covid-19.

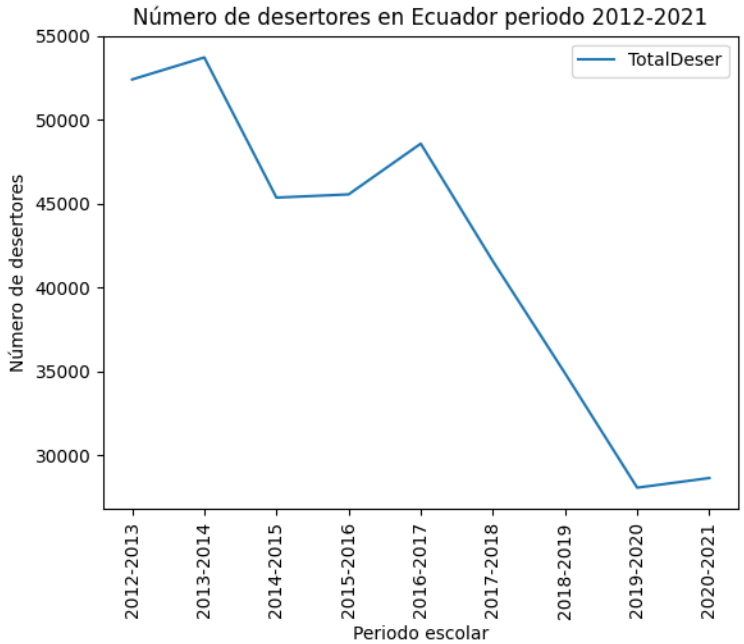


Ilustración 30 Número de estudiantes desertores de bachillerato, adaptado por (Melo 2024)

4.3.2. Creación de dummys para variables categóricas

Luego de haber eliminado algunas variables de nuestro dataset que no se iban a utilizar, vemos cuáles son las variables restantes para transformarlas a dummy. Las variables categóricas por transformar son: Tipo_Educacion, Sostenimiento, Area, Regimen_Escolar, Jurisdiccion, Modalidad, Jornada, TenenciaI, AccesoE. Al generar las variables se eliminó la primera dummy de cada variable para evitar la multicolinealidad en la modelización de los datos.

La ilustración 31 muestra el dataset con la creación de las variables dummy.

	Periodo	Cod_Provincia	DocentesF	DocentesM	DocentesT	AdminF	AdminM	AdminT	PrimeroF	PrimeroM	...	Jornada_Vespertina
0	2012-2013	1	10	7	17	3	0	3	33	25	...	False
1	2012-2013	1	6	11	17	0	0	0	31	38	...	True
2	2012-2013	1	6	2	8	0	0	0	8	14	...	False
3	2012-2013	1	5	7	12	1	2	3	11	10	...	False
4	2012-2013	1	10	9	19	1	0	1	4	7	...	False

5 rows × 98 columns

Ilustración 31 Dataset con dummys, adaptado por (Melo 2024)

Se observa que ahora el dataset contiene 98 variables. Posteriormente se procedió a eliminar las variables de estudiantes de bachillerato por año y género ya que anteriormente se había creado.

La ilustración 32 muestra el dataset con esta reducción de variables, y ahora cuenta con 74 columnas.

	Periodo	Cod_Provincia	DocentesF	DocentesM	DocentesT	AdminF	AdminM	AdminT	TotalBachF	TotalBachM	...	Jornada_Vespertina
0	2012-2013	1	10	7	17	3	0	3	60	50	...	False
1	2012-2013	1	6	11	17	0	0	0	103	93	...	True
2	2012-2013	1	6	2	8	0	0	0	17	29	...	False

3 rows × 74 columns

Ilustración 32 Dataset reducido, adaptado por (Melo 2024)

Para la variable a predecir la cual es TotalDeser, que representa el total de estudiantes que han desertado sus estudios se procedió a transformar esta variable en binario. Es decir, se convirtió a valores de cero cuando no existe ningún estudiante que desertó sus estudios y uno cuando existe uno o más estudiantes que desertaron.

La ilustración 33 muestra la última columna del dataset como la variable de respuesta TotalDeser fue convertida a binaria o booleana.

	TotalBach	TotalProm	TotalPromF	TotalPromM	TotalNoProm	TotalNoPromF	TotalNoPromM	TotalDeser
0	110	88	52	36	0	0	0	1
1	196	144	74	70	27	17	10	1
2	46	33	14	19	3	1	2	1
3	56	52	26	26	0	0	0	1
4	36	35	14	21	1	0	1	0
...	...	...	...	...	...	...	...	...
35805	55	52	20	32	0	0	0	1
35806	25	25	12	13	0	0	0	0
35807	154	152	74	78	0	0	0	1
35808	125	123	43	80	0	0	0	1
35809	7	6	5	1	0	0	0	1

35810 rows × 64 columns

Ilustración 33 Transformación de variable TotalDeser a booleana, adaptado por (Melo 2024)

4.3.3. Reporte de variables

Una vez realizado el preprocesamiento de las variables del dataset, se creó un reporte de las variables para poder observar con detalle las características de cada una. El reporte se adjuntó en el apartado "Anexos".

La ilustración 34 muestra un breve panorama de las variables. Existen 2 variables categóricas Periodo y Cod_provincia, 16 variables numéricas y 56 variables categóricas. Como se observa no existen valores perdidos.

Overview | Alerts 84 | Reproduction

Dataset statistics

Number of variables	74
Number of observations	35810
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	4
Duplicate rows (%)	< 0.1%
Total size in memory	8.8 MiB
Average record size in memory	258.0 B

Variable types

Categorical	2
Numeric	16
Boolean	56

Ilustración 34 Reporte del dataset, adaptado por (Melo 2024)

4.3.4. Exploración estadística

La ilustración 35 muestra las principales estadísticas descriptivas de las variables numéricas del dataset. Vemos que la media de los docentes femenino por escuela es 20.68, para docentes masculino la media es de 10.97, es decir la mitad. La media para las administradoras de género femenino es de 3.48 y 2.56 para el género masculino. La media de estudiantes matriculados en bachillerato de género femenino es de 110.36 y 109.09 para el género masculino. La media de estudiantes promovidos de género femenino es de 98.64 frente a una media de 94.25 estudiantes promovidos de género masculino. La media de estudiantes de género femenino no promovidos es de 2.05 mientras que para el género masculino la media es 3.56 estudiantes.

	Cod_Provincia	DocentesF	DocentesM	DocentesT	AdminF	AdminM	AdminT	TotalBachF	TotalBachM	TotalBach	TotalProm	TotalPromF	TotalPromM
count	35810.000000	35810.000000	35810.000000	35810.000000	35810.000000	35810.000000	35810.000000	35810.000000	35810.000000	35810.000000	35810.000000	35810.000000	35810.000000
mean	11.953588	20.683022	10.978609	31.661631	3.485646	2.562050	6.047696	110.368696	109.093577	219.462273	192.906646	98.646970	94.259676
std	6.669698	21.021944	10.938292	29.703267	6.066024	4.419165	9.883858	172.035326	156.743188	301.021695	269.333767	156.340041	137.835548
min	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	8.000000	7.000000	4.000000	12.000000	1.000000	0.000000	1.000000	23.000000	25.000000	51.000000	41.000000	18.250000	19.000000
50%	12.000000	14.000000	8.000000	21.000000	2.000000	1.000000	3.000000	51.000000	53.000000	109.000000	95.000000	45.000000	45.000000
75%	17.000000	29.000000	14.000000	43.000000	4.000000	3.000000	7.000000	126.000000	126.000000	258.000000	231.000000	114.000000	109.000000
max	90.000000	256.000000	164.000000	357.000000	107.000000	67.000000	170.000000	2875.000000	2712.000000	3649.000000	3380.000000	2633.000000	2497.000000

Ilustración 35 Estadísticas descriptivas, adaptado por (Melo 2024)

La ilustración 36 muestra las clases de la variable de respuesta TotalDeser. El gráfico de dispersión muestra que hay mayor cantidad de estudiantes que desertaron a nivel de escuelas, es decir trabajaremos con clases desbalanceadas.

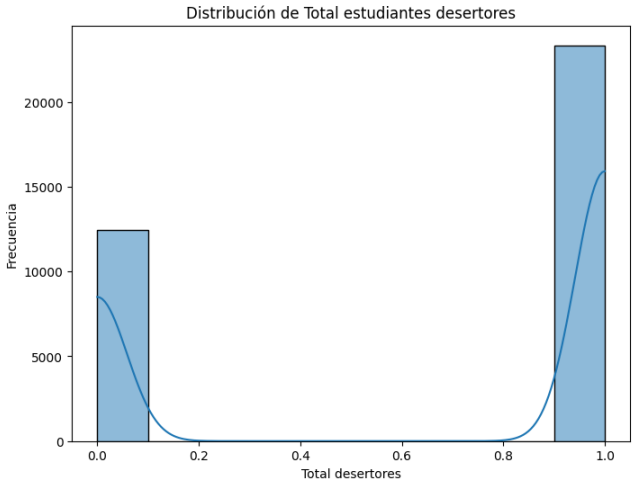


Ilustración 36 Clases de la variable TotalDeser, adaptado por (Melo 2024)

4.3.5. Gráficos de dispersión entre variables predictoras y variable de respuesta

Para analizar la relevancia de las variables predictoras y poder predecir la variable de estudiantes que desertarán sus estudios, se analizará la dispersión ente las variables predictoras y TotalDeser. Ya que nuestra variable de respuesta es binaria o boleana se espera ver dos clases bien diferenciadas:

- 1: indica los estudiantes que han desertado sus estudios.
- 0: indica los estudiantes que no han desertado sus estudios.

La ilustración 37 muestra los valores de la variable DocentesF entre los valores de TotalDeser. Se observa en el gráfico que para valores altos de DocentesF hay más puntos en TotalDeser = 0, por lo tanto, se infiere que mientras exista un mayor número de docentes de género femenino hay menor probabilidad de deserción.

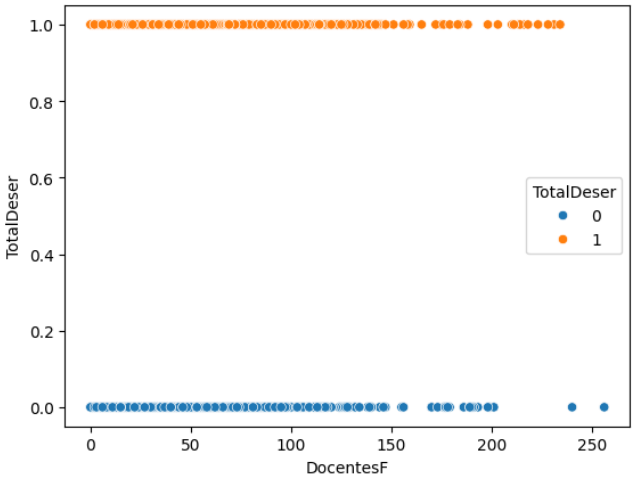


Ilustración 37 Dispersión entre DocentesF y TotalDeser, adaptado por (Melo 2024)

La ilustración 38 muestra los valores de la variable DocentesM entre los valores de TotalDeser. En la gráfica se observa que para valores altos de DocentesM hay más puntos en TotalDeser = 1, es decir se infiere que un mayor número de docentes masculinos está asociado con una mayor probabilidad de deserción escolar.

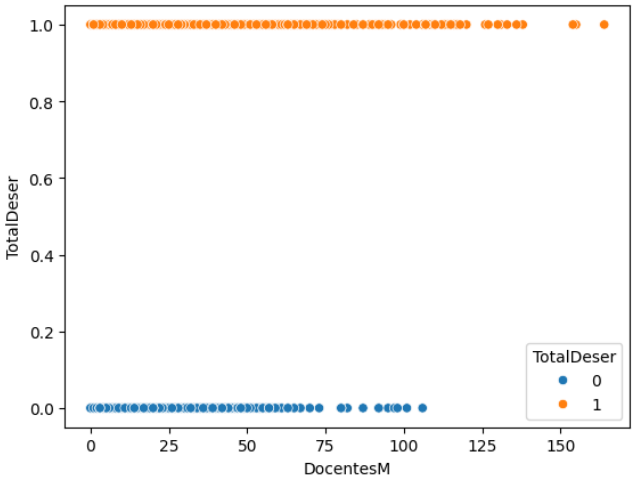


Ilustración 38 Dispersión entre DocentesM y TotalDeser, adaptado por (Melo 2024)

La ilustración 39 muestra los valores de la variable AdminF entre los valores de TotalDeser. Se observa en el gráfico que para valores altos de AdminF y TotalDeser = 1 son similares, por lo tanto, se infiere que esta variable no es un buen predictor de la deserción escolar.

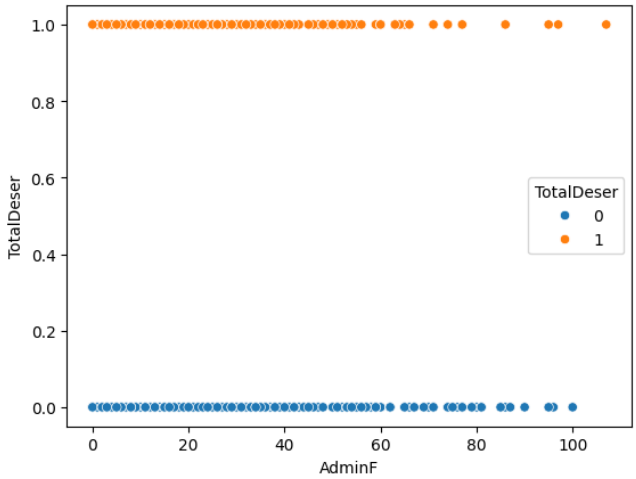


Ilustración 39 Dispersión entre AdminF y TotalDeser, adaptado por (Melo 2024)

La ilustración 40 muestra los valores de la variable AdminM entre los valores de TotalDeser. Se observa en el gráfico que para valores altos de AdminM y TotalDeser = 1 son similares, por lo tanto, se infiere que esta variable no es un buen predictor de la deserción escolar.

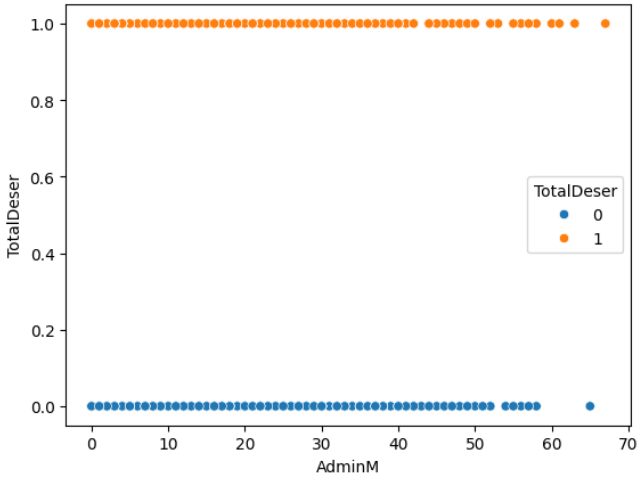


Ilustración 40 Dispersión entre AdminM y TotalDeser, adaptado por (Melo 2024)

La ilustración 41 muestra los valores de la variable TotalBachF entre los valores de TotalDeser. para valores altos de TotalBachF hay más puntos en TotalDeser = 1, es decir se infiere que un mayor número de estudiantes inscritas del género femenino está asociado con una mayor probabilidad de deserción escolar.

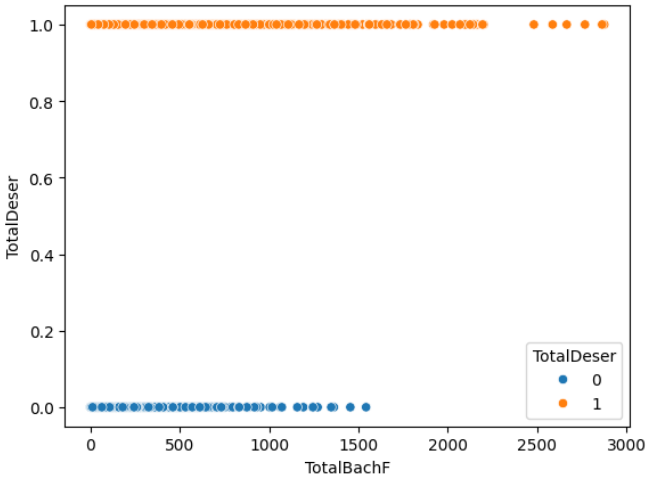


Ilustración 41 Dispersión entre TotalBachF y TotalDeser, adaptado por (Melo 2024)

La ilustración 42 muestra los valores de la variable TotalBachM entre los valores de TotalDeser. Se observa que para valores altos de TotalBachM hay más puntos en TotalDeser = 1, es decir se infiere que un mayor número de estudiantes inscritos de género masculino está asociado con una mayor probabilidad de deserción escolar.

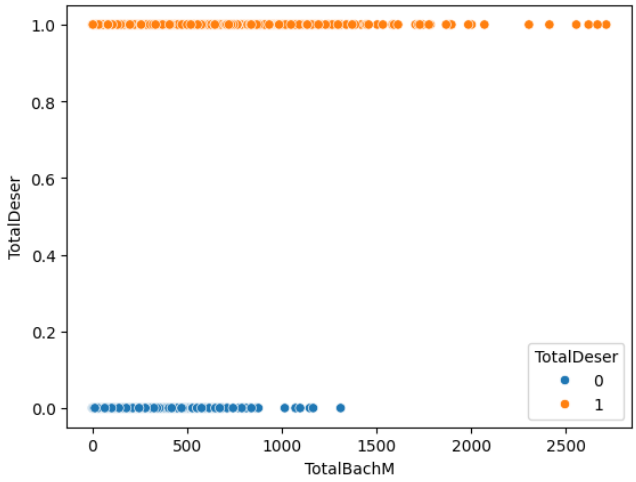


Ilustración 42 Dispersión entre TotalBachM y TotalDeser, adaptado por (Melo 2024)

La ilustración 43 muestra los valores de la variable TotalNoPromF entre los valores de TotalDeser. Se observa que para valores altos de TotalNoPromF hay más puntos en TotalDeser = 1, es decir se infiere que un mayor número de estudiantes no promovidos de género femenino está asociado con una mayor probabilidad de deserción escolar.

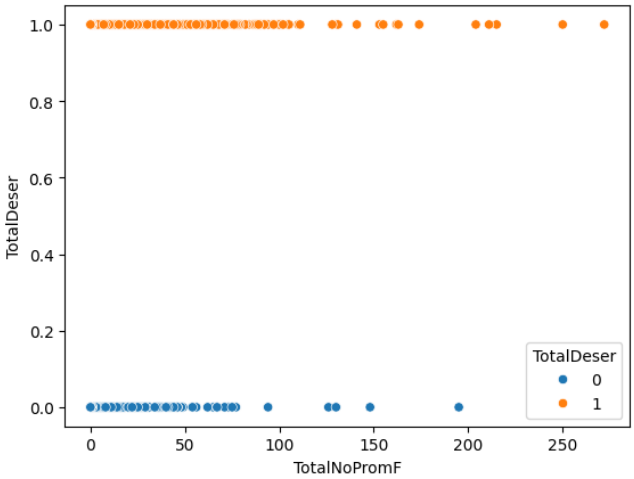


Ilustración 43 Dispersión entre TotalNoPromF y TotalDeser, adaptado por (Melo 2024)

La ilustración 44 muestra los valores de la variable TotalNoPromM entre los valores de TotalDeser. Se observa que para valores altos de TotalNoPromM hay más puntos en TotalDeser = 1, es decir se infiere que un mayor número de estudiantes no promovidos de género masculino está asociado con una mayor probabilidad de deserción escolar.

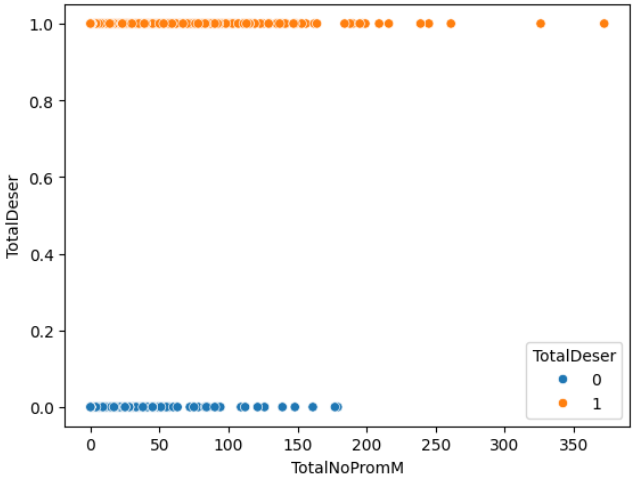


Ilustración 44 Dispersión entre TotalNoPromM y TotalDeser, adaptado por (Melo 2024)

4.3.6. Histogramas

En esta sección se analizó la distribución de las observaciones de las variables predictoras respecto a la variable de respuesta.

La ilustración 45 muestra un histograma entre DocentesF y TotalDeser. La mayoría de las observaciones para ambas categorías de TotalDeser (1 y 0) se encuentran concentradas en los valores más bajos de DocentesF, principalmente entre 0 y 50 docentes de género femenino. Esto indica que a medida que el número de docentes de género femenino aumenta, la frecuencia de ambas categorías disminuye drásticamente.

Existe un solapamiento considerable entre las distribuciones de TotalDeser = 0 y TotalDeser = 1, especialmente en los valores más bajos de DocentesF. Esto sugiere que para valores bajos de DocentesF, es difícil diferenciar entre si los estudiantes desertaron o no, basándose solo en DocentesF.

Las modas (los picos) de ambas distribuciones están bastante cerca, pero la curva naranja (deserción) tiene mayor frecuencia en valores muy bajos de DocentesF en comparación con la curva azul (no deserción). Esto podría indicar que en contextos donde hay menos docentes femeninas, la deserción tiende a ser ligeramente más alta. Esto concuerda con el análisis de dispersiones de esta variable.

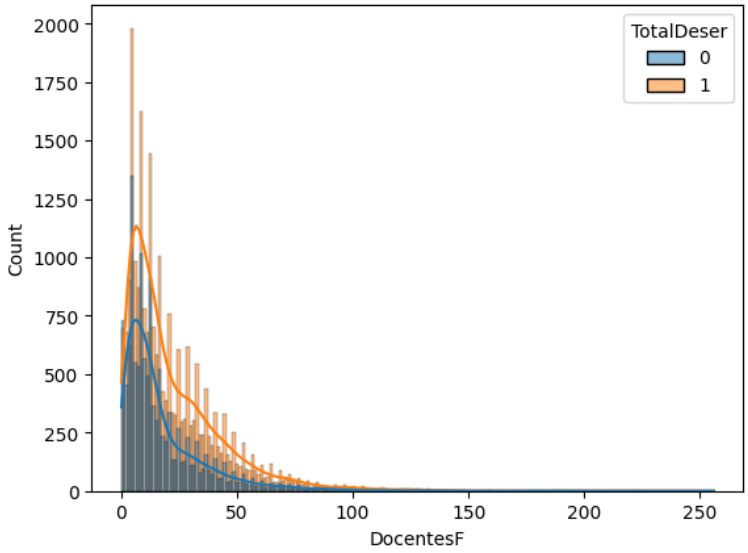


Ilustración 45 Histograma entre DocentesF y TotalDeser, adaptado por (Melo 2024)

La ilustración 46 muestra un histograma entre DocentesM y TotalDeser. Al igual que en el análisis anterior, la mayoría de las observaciones se concentran en los valores más bajos, principalmente entre 0 y 25 docentes de género masculino. Existe un solapamiento considerable entre las distribuciones de TotalDeser = 0 y TotalDeser = 1 y las modas de ambas distribuciones están bastante cercanas, lo que sugiere que es difícil diferenciar entre si los estudiantes desertaron o no, basándose solo en DocentesM.

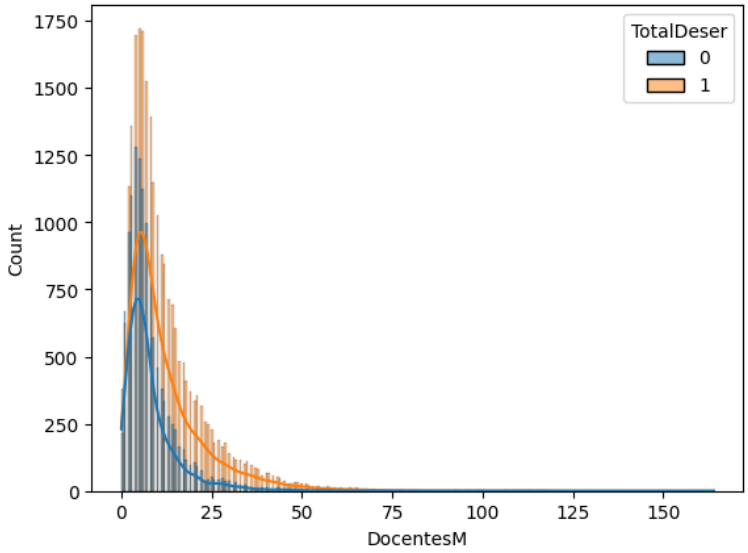


Ilustración 46 Histograma entre DocentesM y TotalDeser, adaptado por (Melo 2024)

Finalmente, se concluye que analizar individualmente estas variables es un buen ejercicio para ver el valor que las variables pueden traer al modelo, sin embargo se debe realizar un análisis conjunto para una mayor precisión.

4.3.7. Matriz de correlación

Con el objetivo de analizar la correlación entre las variables predictoras y la variable de respuesta, se elaboró una matriz de correlación entre estas variables.

La ilustración 47 muestra la matriz de correlación entre las variables.

	TotalDeser	Zona_2	Zona_3	Zona_4	Zona_5	Zona_6	Zona_7	Zona_8	Zona_9	Zona_10
DocentesF	0.107080	-0.031208	-0.034174	-0.066169	-0.000334	-0.064353	-0.050935	0.117960	0.114661	-0.013475
DocentesM	0.188703	-0.000428	0.019454	-0.053874	-0.027096	-0.026623	-0.001351	0.026351	0.062350	-0.008547
DocentesT	0.145274	-0.022244	-0.017022	-0.066669	-0.010214	-0.055348	-0.036545	0.093188	0.104110	-0.012684
AdminF	-0.039094	-0.064099	-0.056812	-0.071351	-0.052610	-0.045026	-0.037232	0.130151	0.214963	-0.017286
AdminM	0.018652	-0.069699	-0.032124	-0.055134	-0.030012	-0.057248	0.000503	0.121194	0.134437	-0.015631
...	...	...	...	...	...	...	...	...	...	...
Tenencial_No conoce	0.002833	-0.009324	-0.010410	-0.000149	-0.021382	0.023184	-0.014539	0.038749	0.005277	-0.004444
Tenencial_Prestado	0.036704	0.046684	-0.022188	0.000929	-0.003779	0.039191	-0.003065	-0.019394	-0.017417	0.020707
Tenencial_Propio	0.046355	-0.015160	0.020379	0.030648	0.055664	0.007821	0.061573	-0.078481	-0.110745	0.006158
AccesoE_Fluvial	-0.020977	0.093748	-0.025419	-0.034420	-0.036870	-0.008670	-0.024336	-0.005642	-0.041324	-0.004391
AccesoE_Terrestre	0.043963	-0.060874	-0.063759	0.045950	0.047604	-0.027912	0.035349	0.021108	0.052260	0.005619

72 rows × 57 columns

Ilustración 47 Matriz de correlación, adaptado por (Melo 2024)

Como se observa, la matriz de correlación tiene muchas columnas por lo que es complicado a simple vista analizar una por una, por ello graficaremos un mapa de calor de las correlaciones para tener una visualización más clara.

La ilustración 48 muestra el mapa de calor de las correlaciones entre las variables, en un rango de -1 a 1.

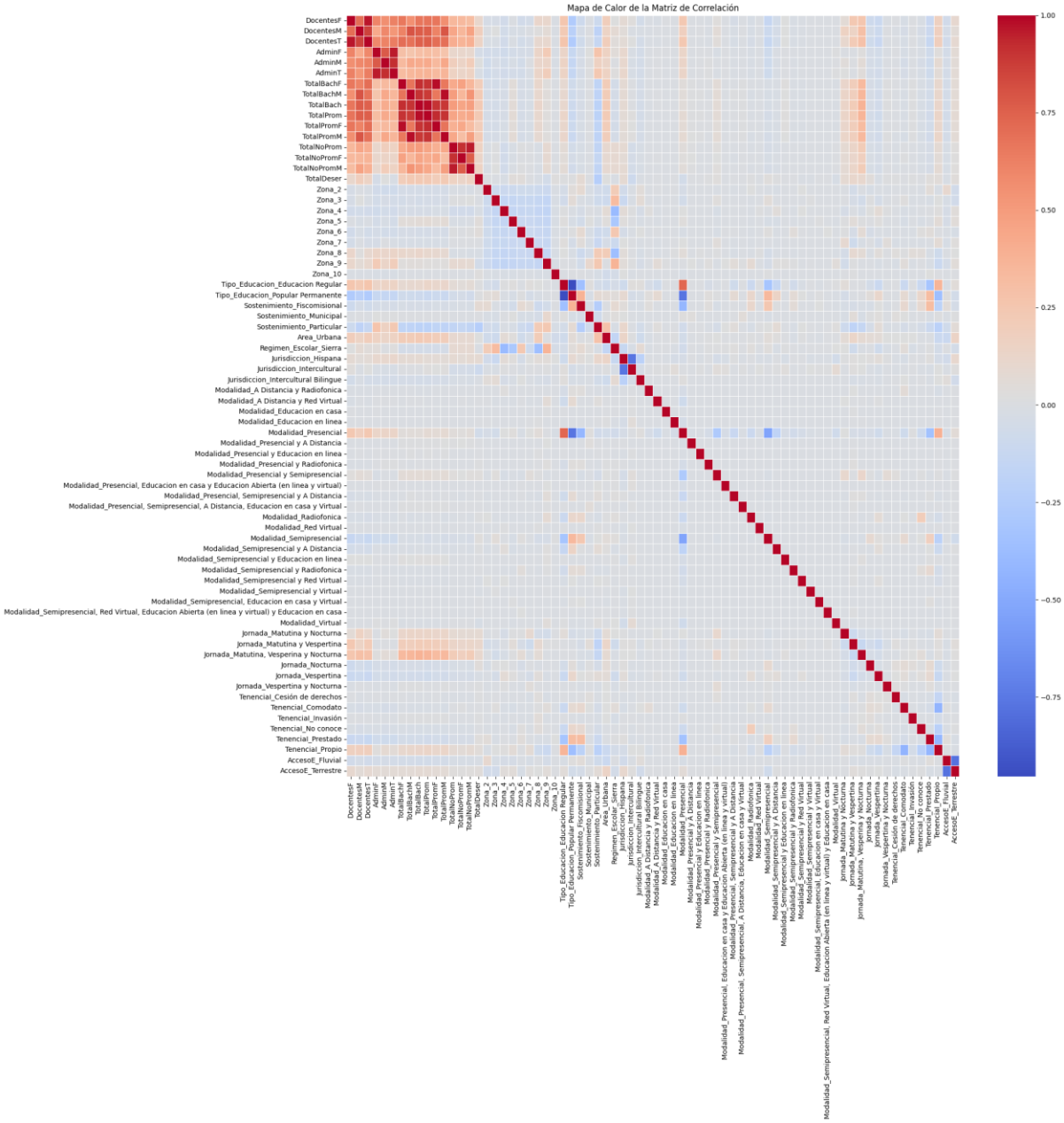


Ilustración 48 Mapa de calor de matriz de correlación, adaptado por (Melo 2024)

La ilustración 49 muestra el mapa de calor de las correlaciones con un filtro para visualizar de mejor manera las correlaciones más altas entre las variables. Se ha resaltado las correlaciones mayores a 0.10.

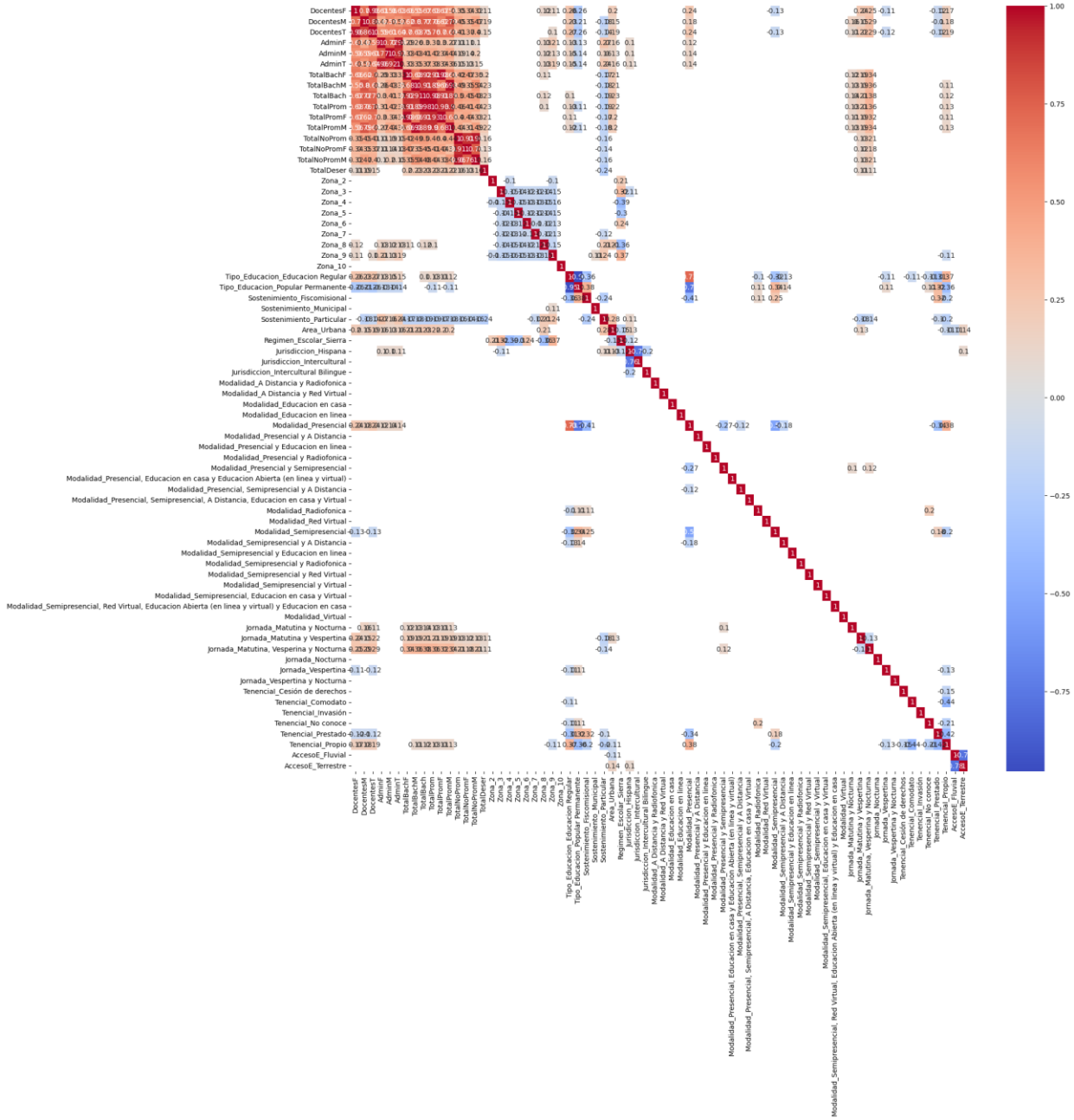


Ilustración 49 Mapa de calor de la matriz de correlación filtrado, adaptado por (Melo 2024)

La ilustración 50 muestra las correlaciones filtradas en la siguiente lista:

TotalDeser	1.000000
TotalBach	0.234341
TotalProm	0.234060
TotalBachM	0.226419
TotalPromM	0.223660
TotalPromF	0.206037
TotalBachF	0.203749
DocentesM	0.188703
TotalNoPromM	0.157056
TotalNoProm	0.155915
DocentesT	0.145274
TotalNoPromF	0.131793
Jornada_Matutina y Vespertina	0.110140
DocentesF	0.107080
Jornada_Matutina, Vesperina y Nocturna	0.105389
Sostenimiento_Particular	-0.235871

Ilustración 50 Correlación de variables con TotalDeser, adaptado por (Melo 2024)

Podemos observar que las variables con la correlación más alta con TotalDeser son TotalBach, TotalProm, TotalBachM, TotalPromM, TotalPromF, TotalBachF, DocentesM, TotalNoPromM, TotalNoProm, DocentesT, TotalNoPromF, Jornada_Matutina y Vespertina, DocentesF, Jornada_Matutina, Vespertina y Nocturna.

4.3.8. Análisis VIF

Una vez identificadas las variables con la correlación más alta respecto a la variable de respuesta, se realizará el análisis VIF o Factor de inflación de la varianza, con el fin de calcular el VIF para cada una de las características y eliminar aquellas con un VIF alto (Statsmodels Variance inflation factor, 2023). Por lo general, se considera un VIF alto por encima de 5 a 10, lo que resulta en una multicolinealidad significativa. Este análisis se hará con el fin evitar la correlación entre las variables predictoras. El VIF se interpreta de la siguiente forma:

- VIF = 1: No hay correlación entre la variable y las demás.
- 1 < VIF < 5: Correlación moderada, generalmente aceptable.
- VIF > 5: Correlación alta, puede ser problemático.
- VIF > 10: Correlación muy alta, se considera severa multicolinealidad.

La ilustración 51 muestra los resultados del cálculo del VIF para cada variable.

Variable	VIF
const	363.911025
DocentesF	inf
DocentesM	inf
DocentesT	inf
AdminF	inf
AdminM	inf
AdminT	inf
TotalBachF	inf
TotalBachM	inf
TotalBach	inf
TotalProm	inf
TotalPromF	inf
TotalPromM	inf
TotalNoProm	inf
TotalNoPromF	inf
TotalNoPromM	inf
Zona_2	1.578256
Zona_3	2.142475
Zona_4	2.600620
Zona_5	2.320205
Zona_6	1.850322
Zona_7	1.850275
Zona_8	2.657973
Zona_9	2.484466
Zona 10	1.029656

Ilustración 51 Cálculo del VIF, adaptado por (Melo 2024)

Tipo_Educacion_Educacion Regular	11.629132
Tipo_Educacion_Popular Permanente	13.569919
Sostenimiento_Fiscomisional	1.593231
Sostenimiento_Municipal	1.050105
Sostenimiento_Particular	2.239742
Area_Urbana	1.390489
Regimen_Escolar_Sierra	2.962115
Jurisdiccion_Hispana	3.369879
Jurisdiccion_Intercultural	3.082783
Jurisdiccion_Intercultural Bilingue	1.143005
Modalidad_A Distancia y Radiofonica	1.008139
Modalidad_A Distancia y Red Virtual	1.025635
Modalidad_Educacion en casa	1.002104
Modalidad_Educacion en linea	1.014523
Modalidad_Presencial	5.193689
Modalidad_Presencial y A Distancia	1.005329
Modalidad_Presencial y Educacion en linea	1.002226
Modalidad_Presencial y Radiofonica	1.009695
Modalidad_Presencial y Semipresencial	1.454107
Modalidad_Presencial, Educacion en casa y Educ...	1.008591
Modalidad_Presencial, Semipresencial y A Dista...	1.042953
Modalidad_Presencial, Semipresencial, A Distan...	1.004549
Modalidad_Radiofonica	1.082302
Modalidad_Red Virtual	1.003521
Modalidad_Semipresencial	1.682783
Modalidad_Semipresencial y A Distancia	1.089934

Ilustración 52 Cálculo del VIF, adaptado por (Melo 2024)

El valor inf muestra que estas variables están perfectamente correlacionadas entre sí, esto puede suceder si alguna variable es una combinación lineal exacta de otras variables en el conjunto. Impacto: Este nivel de colinealidad puede afectar la interpretación de los coeficientes del modelo y hacer que los resultados sean inestables o no confiables.

El análisis indica que hay que eliminar las siguientes variables para evitar la multicolinealidad: DocentesF DocentesT AdminM AdminT TotalBachF TotalBach TotalProm TotalPromF TotalNoProm TotalNoPromF Tipo_Educacion_Educacion Regular Tipo_Educacion_Popular Permanente.

4.3.9. Selección de variables para vector de características

De acuerdo al análisis VIF, se debe descartar las variables mencionadas para evitar la multicolinealidad, adicionalmente se considerará solo aquellas variables con correlación mayor o igual a 0.10.

La ilustración 53 muestra el vector de características X de las variables seleccionadas.

	DocentesM	TotalBachM	TotalPromM	TotalNoPromM	Sostenimiento_Particular	Jornada_Matutina y Vespertina	Jornada_Matutina, Vesperina y Nocturna
0	7	50	36	0	False	False	False
1	11	93	70	10	False	False	False
2	2	29	19	2	False	False	False
3	7	27	26	0	False	False	False
4	9	22	21	1	True	False	False

Ilustración 53 Vector de características X, adaptado por (Melo 2024)

La ilustración 54 muestra el vector de características y para el modelado de datos.

	y
0	1
1	1
2	1
3	1
4	0
5	1
6	1
7	0
8	0
9	1
10	1
11	1
12	1
13	1
14	1
15	0
16	1
17	1

Ilustración 54 Vector de características y, adaptado por (Melo 2024)

4.3.10. División del set de datos en set de prueba y entrenamiento (test y train)

La ilustración 55 muestra la importación de librerías para dividir los datos en un set de entrenamiento y un set de prueba. Posteriormente se selecciona como set test el 20% de las observaciones.

```
# dividir X, y en conjuntos de entrenamiento y prueba
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)
```

Ilustración 55 División del set de prueba y entrenamiento, adaptado por (Melo 2024)

La ilustración 56 muestra las dimensiones del set de entrenamiento y el set de datos. El set de entrenamiento consta de 28.648 filas y 7 columnas y el set de prueba consta de 7.162 filas y 7 columnas.

```
# ver las filas y columnas de X_train  X_test
print('Train set:', X_train.shape, y_train.shape)
print('Test set:', X_test.shape, y_test.shape)
```

✓ 0.0s

```
Train set: (28648, 7) (28648,)
Test set: (7162, 7) (7162,)
```

Ilustración 56 Filas y columnas del set de prueba y entrenamiento, adaptado por (Melo 2024)

4.3.11. Feature engineering

La ilustración 57 muestra los tipos de datos que comprenden el set de entrenamiento en X.

```
# ver tipos de datos de X_train
X_train.dtypes
```

✓ 0.0s

```
DocentesM                              int64
TotalBachM                             int64
TotalPromM                             int64
TotalNoPromM                           int64
Sostenimiento_Particular                bool
Jornada_Matutina y Vespertina          bool
Jornada_Matutina, Vesperina y Nocturna bool
```

Ilustración 57 Comprobar tipos de datos del set train, adaptado por (Melo 2024)

La ilustración 58 muestra la comprobación de datos nulos en el set de entrenamiento de X.

```
# comprobar missings en X_train
X_train.isnull().sum()
```
✓ 0.0s

```
DocentesM                                0
TotalBachM                               0
TotalPromM                               0
TotalNoPromM                             0
Sostenimiento_Particular                 0
Jornada_Matutina y Vespertina            0
Jornada_Matutina, Vesperina y Nocturna   0
```

Ilustración 58 Comprobar NaN en set train, adaptado por (Melo 2024)

La ilustración 59 muestra la comprobación de datos nulos en el set de prueba de X.

```
# comprobar missings en X_test
X_test.isnull().sum()
```
✓ 0.0s

```
DocentesM                                0
TotalBachM                               0
TotalPromM                               0
TotalNoPromM                             0
Sostenimiento_Particular                 0
Jornada_Matutina y Vespertina            0
Jornada_Matutina, Vesperina y Nocturna   0
```

Ilustración 59 Comprobar NaN en set test, adaptado por (Melo 2024)

4.3.12. Feature scaling

Parte del proceso antes del modelado de datos se debe normalizar o estandarizar los mismos para manejar escalas iguales, así los datos tendrán una magnitud comparable entre sí.
La ilustración 60 muestra las estadísticas descriptivas de las variables, como podemos ver están en escalas distintas, por lo que se procedió a normalizarlas.

```
X_train.describe()
```
✓ 0.0s

	DocentesM	TotalBachM	TotalPromM	TotalNoPromM
count	28648.000000	28648.000000	28648.000000	28648.000000
mean	10.936854	109.166225	94.327702	3.530718
std	10.821967	156.645793	137.712052	11.231005
min	0.000000	0.000000	0.000000	0.000000
25%	4.000000	24.000000	19.000000	0.000000
50%	8.000000	53.000000	45.000000	0.000000
75%	14.000000	126.000000	109.000000	2.000000
max	136.000000	2712.000000	2497.000000	372.000000

Ilustración 60 Estadísticas del set test, adaptado por (Melo 2024)

La ilustración 61 muestra el proceso de normalización de los datos para el set de test y train respectivamente.

```
# normalización de los datos
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()

X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

Ilustración 61 Normalización del set train y test, adaptado por (Melo 2024)

La ilustración 63 muestra la normalización del set train y sus estadísticas principales.

```
X_train.describe()
```
✓ 0.0s Python

	DocentesM	TotalBachM	TotalPromM	TotalNoPromM	Sostenimiento_Particular	Jornada_Matutina y Vespertina	Jornada_Matutina, Vesperina y Nocturna
count	28648.000000	28648.000000	28648.000000	28648.000000	28648.000000	28648.000000	28648.000000
mean	0.080418	0.040253	0.037776	0.009491	0.331576	0.215931	0.056583
std	0.079573	0.057760	0.055151	0.030191	0.470788	0.411474	0.231049
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.029412	0.008850	0.007609	0.000000	0.000000	0.000000	0.000000
50%	0.058824	0.019543	0.018022	0.000000	0.000000	0.000000	0.000000
75%	0.102941	0.046460	0.043652	0.005376	1.000000	0.000000	0.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Ilustración 62 Comprobación de la normalización del set train, adaptado por (Melo 2024)

4.4. Etapa de minería de datos

4.4.1. Regresión logística

Para el modelado de la regresión logística se importaron las librerías necesarias y se ajustó el hiperparámetro de la regresión logística C = 1. La ilustración 63 muestra el entrenamiento del modelo de regresión logística en el conjunto de datos de entrenamiento (train).

```
# importar librerías
from sklearn.linear_model import LogisticRegression

# entrenar un modelo de regresión logística en el conjunto de entrenamiento
LR= LogisticRegression(C = 1, solver='liblinear', random_state=0).fit(X_train, y_train) # ajustar el modelo con fit
LR
```
✓ 0.1s

```
LogisticRegression
LogisticRegression(C=1, random_state=0, solver='liblinear')
```

Ilustración 63 Entrenamiento de regresión logística, adaptado por (Melo 2024)

La ilustración 64 muestra las predicciones de los resultados en el set de entrenamiento, donde 0 indica la probabilidad de no desertar y 1 indica la probabilidad de desertar. El resultado es un arreglo de valores entre 0 y 1.

```
# Predecir los resultados con set test
y_hat = LR.predict(X_test)
y_hat
```
✓ 0.0s Open 'y_hat' in Data Wrangler

```
array([1, 0, 0, ..., 0, 1, 1], dtype=int64)
```

Ilustración 64 Predicción de datos del set test en regresión logística, adaptado por (Melo 2024)

La ilustración 65 muestra la probabilidad de obtener la salida 0, no desertor. La función predict_proba devuelve estimaciones para todas las clases, ordenadas por la etiqueta de las clases. La primera columna corresponde a la probabilidad clase 0, mientras que la segunda columna es la probabilidad de la clase 1.

```
# probabilidad de obtener salida como 0 (no desertor)
from sklearn.metrics import precision_recall_curve
y_hat_prob_0 = LR.predict_proba(X_test)
y_hat_prob_0
```
✓ 0.0s Open 'y_hat_prob_0' in Data Wrangler

```
array([[0.49267916, 0.50732084],
       [0.55532676, 0.44467324],
       [0.53071073, 0.46928927],
       ...,
       [0.51886351, 0.48113649],
       [0.35556012, 0.64443988],
       [0.4957171 , 0.5042829 ]])
```

Ilustración 65 Predicción de no desertor en regresión logística, adaptado por (Melo 2024)

La ilustración 66 muestra la probabilidad de obtener la salida 1 que representa a los desertores en el conjunto de datos de entrenamiento.

```
# probabilidad de obtener salida como 1 = desertar
y_hat_prob_1 = LR.predict_proba(X_test)[:,1]
y_hat_prob_1
```
✓ 0.0s Open 'y_hat_prob_1' in Data Wrangler

```
array([0.50732084, 0.44467324, 0.46928927, ..., 0.48113649, 0.64443988,
       0.5042829 ])
```

Ilustración 66 Predicción de desertor en regresión logística, adaptado por (Melo 2024)

4.4.2. Decision Trees o árboles de decisión

Los árboles de decisión o decisión trees también permiten hacer predicciones de categorías. Estos modelos de clasificación se construyen dividiendo los datos de entrenamiento en distintos nodos, donde un nodo contiene todas o la mayor parte de una categoría de los datos. luego de evaluar un atributo, ramifican los casos basados en los resultados de la evaluación.

La ilustración 67 muestra el entrenamiento del modelo decision tree con los sets de datos previamente determinados. Con la función y_pred se predicen los valores para la variable de respuesta TotalDeser. El resultado refleja un arreglo de valores de 0 y 1 de acuerdo a las predicciones realizadas con el algoritmo.

```
# importar librerías
from sklearn.tree import DecisionTreeClassifier

# entrenar el modelo
Classifier = DecisionTreeClassifier()
Classifier.fit(X_train, y_train)

# predecir set test
y_hat_DT = Classifier.predict(X_test)
print(y_hat_DT)
```

```
[1 1 1 ... 0 0 0]
```

Ilustración 67 Entrenamiento de modelo decision tree, adaptado por (Melo 2024)

Posteriormente se realizó una representación textual del árbol de decisión entrenado, donde el árbol comienza por la raíz o el primer nodo. La ilustración 86 muestra una visualización del árbol de decisión.

```
# visualizar el árbol de decisión
from sklearn import tree
text_representation = tree.export_text(Classifier)
print(text_representation)
```

```
|--- feature_2 <= 0.00
|   |--- feature_1 <= 0.00
|   |   |--- feature_0 <= 0.06
|   |   |   |--- feature_6 <= 0.50
|   |   |   |   |--- feature_2 <= 0.00
|   |   |   |   |   |--- feature_5 <= 0.50
|   |   |   |   |   |   |--- feature_1 <= 0.00
|   |   |   |   |   |   |   |--- feature_0 <= 0.03
|   |   |   |   |   |   |   |   |--- feature_0 <= 0.00
|   |   |   |   |   |   |   |   |   |--- feature_1 <= 0.00
|   |   |   |   |   |   |   |   |   |   |--- feature_4 <= 0.50
|   |   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 2
|   |   |   |   |   |   |   |   |   |   |--- feature_4 >  0.50
|   |   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 2
|   |   |   |   |   |   |   |   |   |--- feature_1 >  0.00
|   |   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |--- feature_0 >  0.00
|   |   |   |   |   |   |   |   |   |--- feature_1 <= 0.00
|   |   |   |   |   |   |   |   |   |   |--- feature_0 <= 0.01
|   |   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 3
|   |   |   |   |   |   |   |   |   |--- feature_0 >  0.01
|   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 5
```

Ilustración 68 Visualización de decision tree, adaptado por (Melo 2024)

La ilustración 69 muestra la visualización del árbol de decisión para ver el proceso de creación de nodos.

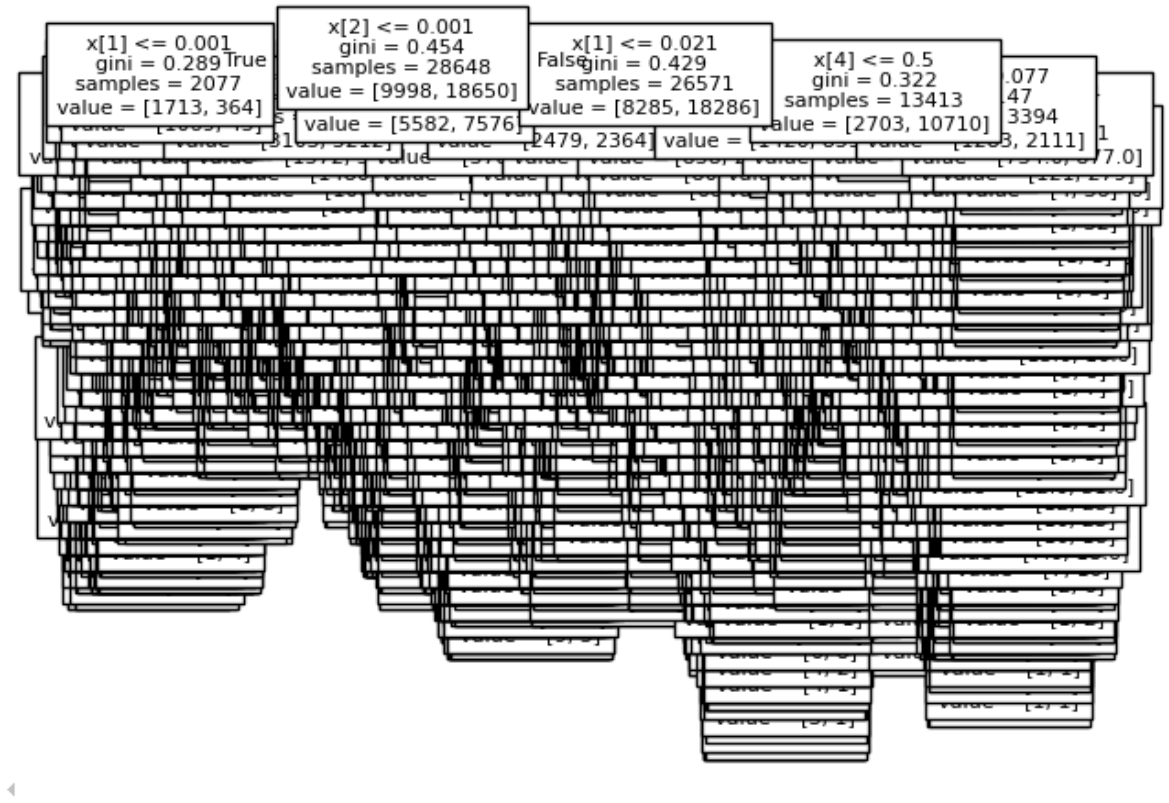


Ilustración 69 Visualización dos de decision tree, adaptado por (Melo 2024)

La ilustración 70 muestra el árbol de decisión con las clases. En el árbol de decisión vemos que las hojas azules son las clases positivas y las naranjas las clases negativas, y el árbol va creando las hojas de acuerdo al valor de la variable que se le asignó.

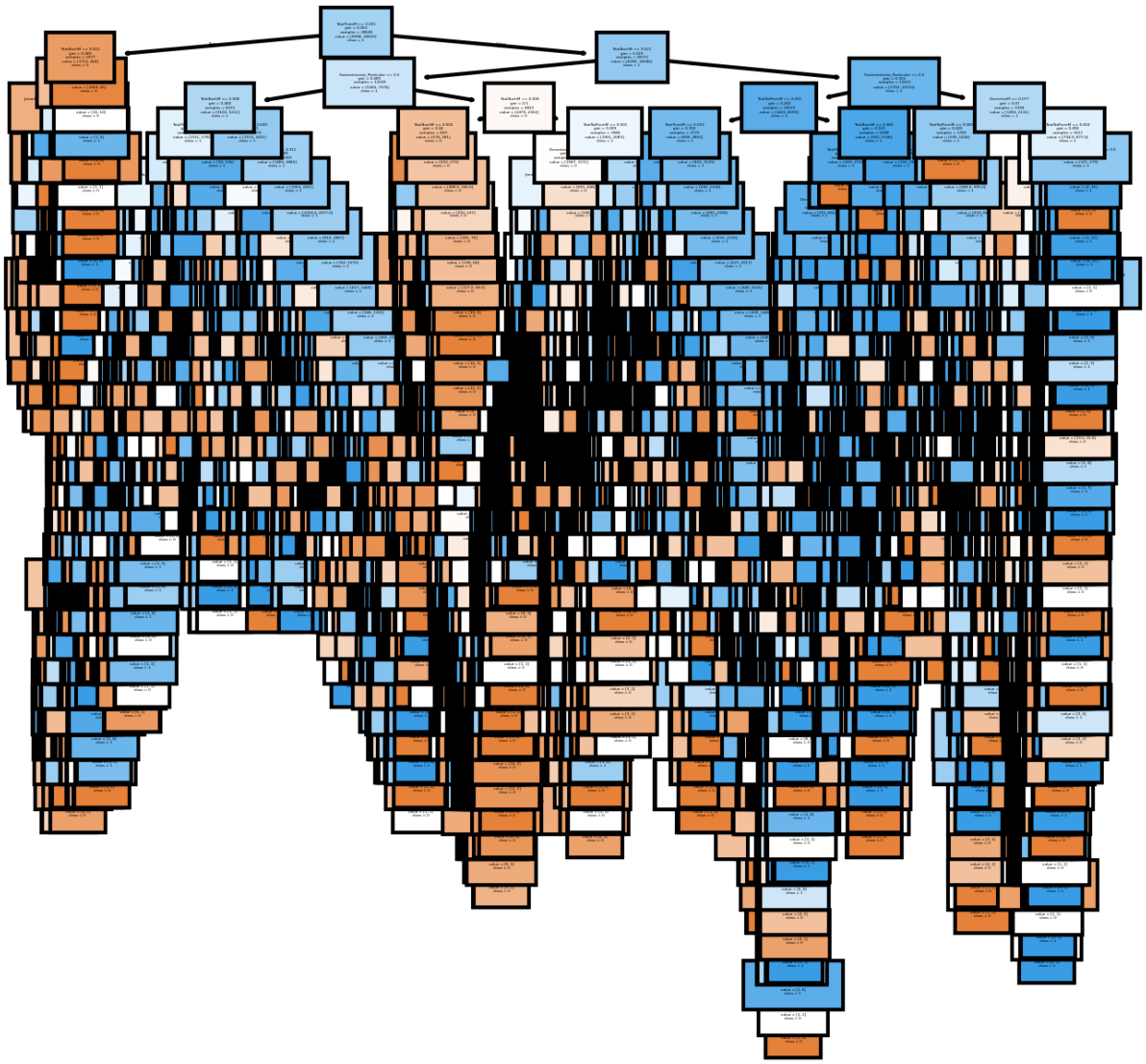


Ilustración 70 Visualización tres de decision tree, adaptado por (Melo 2024)

4.4.3. Importancia de las características

Se extrajo la importancia de las características del modelo de árbol de decisión. La ilustración 71 muestra las características y la importancia de cada una. La más importante para el modelo es TotalPromM con 0.3510, esta característica representa el total de estudiantes promovidos de año escolar de género masculino.

```python
# extraer la importancia de las features o características
feature_names = X.columns
feature_importances = Classifier.feature_importances_

# imprimir la importancia de las features
for name, importance in zip(feature_names, feature_importances):
    print(f"{name}: {importance:.4f}")
```

✓ 0.0s

```
DocentesM: 0.1627
TotalBachM: 0.3340
TotalPromM: 0.3510
TotalNoPromM: 0.0817
Sostenimiento_Particular: 0.0374
Jornada_Matutina y Vespertina: 0.0260
Jornada_Matutina, Vesperina y Nocturna: 0.0072
```

Ilustración 71 Características o features de decision tree, adaptado por (Melo 2024)

4.4.4. Random Forest

el modelo de clasificación ramdom forest es una versión mejorada del árbol de decisión. Usaremos el mismo set de datos que los modelos anteriores. Esta función también tiene un hiperparámetro para determinar cuántos árboles de decisión tendrá el bosque aleatorio, el usualmente recomendado es 100, pero en este caso se generó 20, ya que con 100 el accuracy era mucho menor.

La ilustración 72 muestra el entrenamiento del modelo random forest para el mismo set de datos de prueba.

```python
# ajustar modelo
from sklearn.ensemble import RandomForestRegressor

Classifier2 = RandomForestRegressor(n_estimators=20, random_state=42)
Classifier2.fit(X_train, y_train)

# predecir test set
y_hat_RF = Classifier2.predict(X_test)
predictions = [round(value) for value in y_hat_RF]
```

Ilustración 72 Entrenamiento del modelo random forest, adaptado por (Melo 2024)

4.4.5. K-Nearest Neighbors o K-vecinos más cercanos

El algoritmo K- Nearest Neighbors clasifica los nuevos puntos de datos según la similitud. El hiperparámetro número de vecinos se puede configurar, siendo el default = 5, por lo general se usa un número impar de vecinos.

```
# importar librerías
from sklearn.neighbors import KNeighborsClassifier

# ajustar el modelo
Classifier3 = KNeighborsClassifier(n_neighbors = 5)
Classifier3.fit(X_train, y_train)

# predecir set test
y_hat_KNN = Classifier3.predict(X_test)
predictions = [round(value) for value in y_hat_KNN]
```

Ilustración 73 Entrenamiento del modelo K-Nearest Neighbors, adaptado por (Melo 2024)

4.5.6. Support Vector Machine (SVM)

Este modelo de clasificación es muy útil cuando se tienen datos que no son fáciles de identificar, este modelo crea un plano de división entre los datos y los agrupa según sus características. Importamos la función LinearSVC y en este caso el hiperparámetro del modelo es lineal, el cual permite crear un hiperplano entre las variables. Hay otros hiperparámetros que se pueden usar dependiendo de la naturaleza del problema.

La ilustración 74 muestra el entrenamiento del modelo SVM.

```
# importar librerías
from sklearn import svm
Classifier4 = svm.LinearSVC(random_state = 20)

# ajustar el modelo
Classifier4.fit(X_train, y_train)

# predecir set test
y_hat_SVM = Classifier4.predict(X_test)
```

Ilustración 74 Entrenamiento del modelo SVM, adaptado por (Melo 2024)

4.5. Etapa de interpretación y evaluación de datos

4.5.1. Evaluación de la regresión logística

4.5.1.1. Accuracy

Accuracy es una métrica de evaluación que mide la predicción correcta de las clases, es ampliamente usado para evaluar el rendimiento de un modelo de clasificación. Accuracy se calcula como la proporción de predicciones correctas respecto al total de predicciones realizadas.

La ilustración 75 muestra el cálculo de accuracy para el modelo de regresión logística. En este caso, es de 67,36%.

```
from sklearn.metrics import accuracy_score
accuracy_LR = accuracy_score(y_test, y_hat)
print('Accuracy score del modelo: {0:0.4f}',accuracy_LR)
```

```
Accuracy score del modelo: {0:0.4f} 0.6736944987433677
```

Ilustración 75 Accuracy de la regresión logística, adaptado por (Melo 2024)

4.5.1.2. Revisión de overfitting y underfitting

Es necesario comprobar uno de los problemas más comunes en los algoritmos de machine learning. El overfitting sucede cuando un modelo es demasiado complejo y se ajusta muy bien a los datos de entrenamiento, capturando ruido y patrones aleatorios. Esto resulta en un buen rendimiento en el conjunto de entrenamiento, pero un pobre desempeño en nuevos datos (conjunto de prueba) porque no generaliza bien el modelo, ya que el modelo es muy flexible y se ajusta más al ruido antes que la función que buscamos.

El underfitting sucede cuando un modelo es demasiado simple para capturar la estructura subyacente de los datos. Puede que no aprenda lo suficiente de los datos de entrenamiento, lo que lleva a un bajo rendimiento tanto en el conjunto de entrenamiento como en el conjunto de prueba. Si incrementamos el orden del polinomio, el modelo se ajusta mejor, pero el modelo no es lo suficientemente flexible y presenta underfitting.

La ilustración 76 muestra la revisión del overfitting y underfitting del modelo de regresión logística. Esto indica que el modelo tiene una precisión de aproximadamente el 66.88% en los datos de entrenamiento y del 66.02% en los datos de prueba.

Esto sugiere que el modelo está rindiendo de manera similar en ambos conjuntos de datos, lo que generalmente es una señal de que no está sobreajustado (overfitted) ni subajustado (underfitted).

```
# Revisar overfitting y underfitting

print('Training set score: {:.4f}'.format(LR.score(X_train, y_train)))
print('Test set score: {:.4f}'.format(LR.score(X_test, y_test)))
```

```
Training set score: 0.6807
Test set score: 0.6737
```

Ilustración 76 Revisión de overfitting y underfitting en regresión logística, adaptado por (Melo 2024)

4.5.1.3. Matriz de confusión

Para evaluar las predicciones del modelo se graficó la matriz de confusión, con el fin de tener una visión más completa al evaluar el rendimiento de un modelo. Para analizar la matriz de confusión se debe revisar los casos en la diagonal principal de la matriz. En este estudio nos interesa identificar los estudiantes que desertarán, por lo tanto los verdaderos positivos (TP).
la ilustración 77 muestra la matriz de confusión para el modelo de regresión logística.

Análisis:

- La primera fila corresponde a TotalDeser = 1 en el conjunto de prueba = 1 (TP), es decir del total de datos 4.693 el modelo predijo correctamente 3.901 casos como estudiantes que desertarán (TotalDeser = 1). Los falsos negativos (FN) indican que el modelo predijo 792 casos como 0 (TotalDeser = 0) cuando la etiqueta correcta era (TotalDeser = 1).
- La segunda fila corresponde a los falsos positivos (FP) lo que indica que de 2.469 casos, el modelo predijo 1.545 con la etiqueta (TotalDeset = 1) cuando la etiqueta correcta era (TotalDeser =0) y 924 casos se predijeron correctamente como verdaderos negativos (TN).

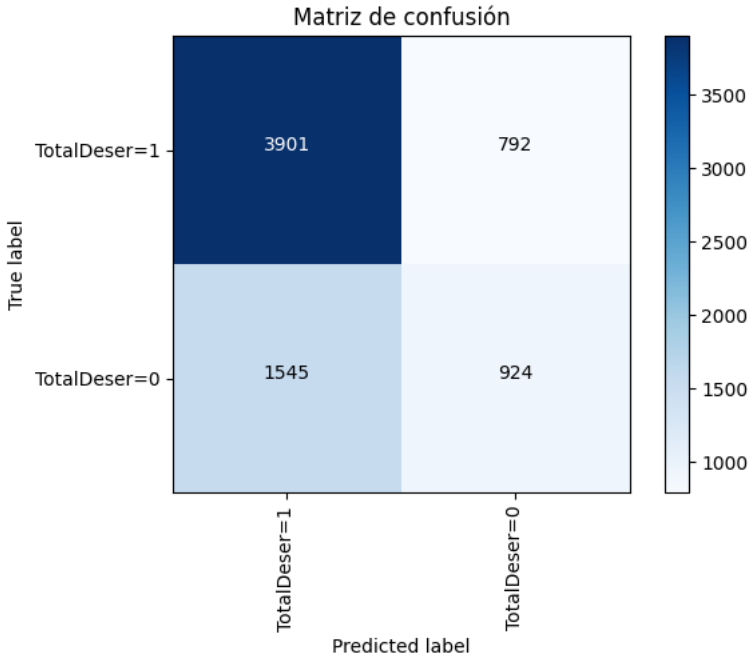


Ilustración 77 Matriz de confusión de la regresión logística, adaptado por (Melo 2024)

4.5.1.4. Otras métricas de evaluación

Así como el accuracy, existen otras métricas para evaluar el modelo de clasificación, dependiendo del interés del investigador en predecir cierta etiqueta, en este caso los verdaderos positivos que corresponde a la probabilidad de que un estudiante desierte en sus estudios.
La ilustración 78 muestra una tabla que resume las métricas de evaluación precision, recall, f1-score y support.

Análisis:

- **Precision:** es una medida de la exactitud siempre que se haya predicho una etiqueta de clase. Se define por: precisión = TP / (TP + FP). En este caso nos interesa evaluar la precisión ya que se usa para medir la exactitud de la clase 1 (TotalDeser = 1). En este modelo la precisión es 72% para la clase 1.
- **Recall:** el recall es del 83%. Esto significa que el 83% de todos los ejemplos que realmente pertenecen a la clase 1 fueron correctamente identificados por el modelo.
- **F1-score:** es el promedio armónico de precision y accuracy, donde una puntuación F1 alcanza su mejor valor en 1 (precisión y recuperación perfectas) y su peor valor en 0. Es una buena forma de demostrar que un clasificador tiene un buen valor para ambos, recall y precision y es ampliamente usado en clases desbalanceadas, como en este caso. El F1-score para este modelo es 77% para la clase (TotalDeser = 1), indicando un buen balance entre precisión y recall para la clase 1.
- **Support:** para la clase 1 hay 4693 ejemplos de la clase 1 en el conjunto de prueba.
- **Accuracy:** el accuracy general del modelo es del 67%, lo que significa que el 67% de todas las predicciones (tanto clase 0 como clase 1) fueron correctas.

	precision	recall	f1-score	support
0	0.54	0.37	0.44	2469
1	0.72	0.83	0.77	4693
accuracy			0.67	7162
macro avg	0.63	0.60	0.61	7162
weighted avg	0.65	0.67	0.66	7162

Ilustración 78 Otras métricas de evaluación para regresión logística, adaptado por (Melo 2024)

4.5.1.5. Log loss o pérdida logística

La pérdida logística o log loss es otra métrica que sirve para evaluar el modelo. En la regresión logística, la salida puede ser la probabilidad de que la pérdida de clientes sea sí (o igual a 1). Esta probabilidad es un valor entre 0 y 1.

La ilustración 79 muestra el cálculo de log loss para el modelo de regresión logística. En este caso log loss equivale a 59,13%.

```
# log loss o pérdida logística
from sklearn.metrics import log_loss
log_loss(y_test, y_hat_prob_1)
```

```
0.5913130374485953
```

Ilustración 79 Pérdida logística de regresión logística, adaptado por (Melo 2024)

4.5.1.6. Curva ROC y curva PR

Otra herramienta para medir visualmente el rendimiento del modelo de clasificación es la curva ROC. Curva ROC son las siglas de Receiver Operating Characteristic Curve. Una curva ROC es un gráfico que muestra el rendimiento de un modelo de clasificación en varios niveles de umbral de clasificación. La curva ROC representa la tasa de verdaderos positivos (TPR) frente a la tasa de falsos positivos (FPR) en varios niveles de umbral.

La curva PR ayuda a medir visualmente las métricas precision y recall. Esta gráfica permite determinar a partir de qué recall se degrada precision y viceversa. El ideal de esta curva es que se acerque lo más posible a la esquina superior derecha, lo que representaría alto precision y alto recall.

La ilustración 80 muestra las curvas ROC y PR.

Análisis:

- **Curva ROC:** la curva ROC consta de dos ejes, el eje X representa la tasa de falsos positivos y el eje y representa los verdaderos positivos, el área bajo la curva mide el rendimiento global del modelo, en este caso el AUC = 0.71 que señala un buen rendimiento del modelo. Un AUC = 1 indica un modelo perfecto.
- **Curva PR:** un PR AUC = 0.81 sugiere que el modelo tiene un rendimiento razonablemente bueno en términos de precisión y recall. No es perfecto, pero muestra una capacidad significativa para distinguir entre las clases positivas y negativas.

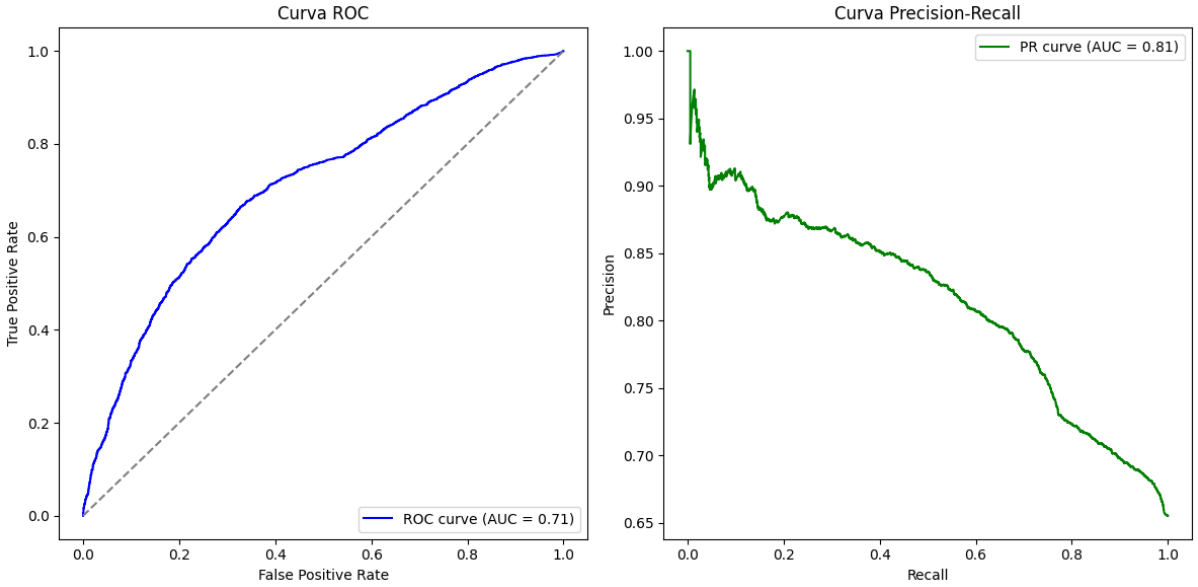


Ilustración 80 Curvas ROC y PR de regresión logística, adaptado por (Melo 2024)

4.5.1.7. Cálculo de residuos

Se calcularon los residuos del modelo, utilizando el valor observado menos el valor predicho para determinar si existen Nan en las predicciones, del set de datos de prueba. La ilustración 81 muestra los resultados de este cálculo. Se determina que no existen NaN en los residuos.

```
# Verificar si hay NaN o Inf en las predicciones
print(np.isnan(y_hat).sum(), np.isinf(y_hat).sum())

# Verificar si hay NaN o Inf en los residuos
print(np.isnan(residuals).sum(), np.isinf(residuals).sum())
```

```
0 0
0 0
```

Ilustración 81 Verificación de residuos de regresión logística, adaptado por (Melo 2024)

4.5.2. Evaluación de decision trees o árboles de decisión.

4.5.2.1. Matriz de confusión

La evaluación de las predicciones con el árbol de decisión se realizó a través de la matriz de confusión, como en el caso del modelo anterior. La ilustración 82 muestra la matriz de confusión.

Análisis:

- La diagonal principal muestra que los verdaderos positivos predichos corresponde a 3.827 casos, mientras que 866 casos se han predicho incorrectamente como positivos, siendo negativos.
- Los verdaderos negativos corresponden a 1.769 casos que se han predicho correctamente, mientras que 700 casos se han predicho incorrectamente como negativos, siendo de clase positiva.
- Se ha calculado la métrica accuracy para este modelo con un 78,13% de precisión en las predicciones.

```
# Evaluar las predicciones con accuracy
cm = confusion_matrix(y_test, y_hat_DT)
print(cm)
accuracy_DT = accuracy_score(y_test,y_hat_DT)
print("Accuracy: %.2f%%" % (accuracy_DT * 100.0))
```

✓ 0.0s

```
[[1769  700]
 [ 866 3827]]
Accuracy: 78.13%
```

Ilustración 82 Matriz de confusión de decision tree, adaptado por (Melo 2024)

4.5.2.2. Otras métricas de evaluación

La ilustración 83 muestra la tabla de métricas de evaluación para el modelo de árbol de decisión.

Análisis:

- **Precision:** en este caso nos interesa evaluar la precisión ya que se usa para medir la exactitud de la clase 1 (TotalDeser = 1). En este modelo la precisión es 85% para la clase 1.
- **Recall:** el recall es del 82%. Esto significa que el 82% de todos los ejemplos que realmente pertenecen a la clase 1 fueron correctamente identificados por el modelo.
- **F1-score:** es el promedio armónico de precision y accuracy, donde una puntuación F1 alcanza su mejor valor en 1 (precisión y recuperación perfectas) y su peor valor en 0. El F1-score para este modelo es 83% para la clase (TotalDeser = 1), indicando un buen balance entre precisión y recall para la clase 1.
- **Support:** para la clase 1 hay 4693 ejemplos de la clase 1 en el conjunto de prueba.
- **Accuracy:** el accuracy general del modelo es del 78%, lo que significa que el 78% de todas las predicciones (tanto clase 0 como clase 1) fueron correctas.

Reporte de Clasificación de DT:

	precision	recall	f1-score	support
0	0.67	0.72	0.69	2469
1	0.85	0.82	0.83	4693
accuracy			0.78	7162
macro avg	0.76	0.77	0.76	7162
weighted avg	0.79	0.78	0.78	7162

Ilustración 83 Otras métricas de evaluación para decision tree, adaptado por (Melo 2024)

4.5.3. Evaluación de Random Forest

4.5.3.1. Matriz de confusión

Se elaboró la matriz de confusión para el modelo random forest. La ilustración 84 muestra la matriz de confusión.

Análisis:

- La diagonal principal muestra que los verdaderos positivos predichos corresponde a 4.066 casos, mientras que 627 casos se han predicho incorrectamente como positivos, siendo negativos.
- Los verdaderos negativos corresponden a 1.758 casos que se han predicho correctamente, mientras que 711 casos se han predicho incorrectamente como negativos, siendo de clase positiva.
- Se ha calculado la métrica accuracy para este modelo con un 81,32% de precisión en las predicciones.

```
# evaluar las predicciones con accuracy
cm2 = confusion_matrix(y_test, predictions)
print(cm2)
accuracy_RF = accuracy_score(y_test,predictions)
print("Accuracy: %.2f%%" % (accuracy_RF * 100.0))
```

✓ 0.0s

```
[[1758  711]
 [ 627 4066]]
Accuracy: 81.32%
```

Ilustración 84 Matriz de confusión de random forest, adaptado por (Melo 2024)

4.5.3.2. Otras métricas de evaluación

La ilustración 85 muestra la tabla de métricas de evaluación para el modelo de random forest.

Análisis:

- **Precision:** en este caso nos interesa evaluar la precisión ya que se usa para medir la exactitud de la clase 1 (TotalDeser = 1). En este modelo la precisión es 85% para la clase 1.
- **Recall:** el recall es del 87%. Esto significa que el 87% de todos los ejemplos que realmente pertenecen a la clase 1 fueron correctamente identificados por el modelo.
- **F1-score:** es el promedio armónico de precision y accuracy, donde una puntuación F1 alcanza su mejor valor en 1 (precisión y recuperación perfectas) y su peor valor en 0. El F1-score para este modelo es 86% para la clase (TotalDeser = 1), indicando un buen balance entre precisión y recall para la clase 1.
- **Support:** para la clase 1 hay 4693 ejemplos de la clase 1 en el conjunto de prueba.
- **Accuracy:** el accuracy general del modelo es del 81%, lo que significa que el 81% de todas las predicciones (tanto clase 0 como clase 1) fueron correctas.

```
Reporte de Clasificación de RF:
              precision    recall  f1-score   support

           0       0.74      0.71      0.72      2469
           1       0.85      0.87      0.86      4693

    accuracy                           0.81      7162
   macro avg       0.79      0.79      0.79      7162
weighted avg       0.81      0.81      0.81      7162
```

Ilustración 85 Otras métricas de evaluación para random forest, adaptado por (Melo 2024)

4.5.4. Evaluación de K-Nearest Neighbors

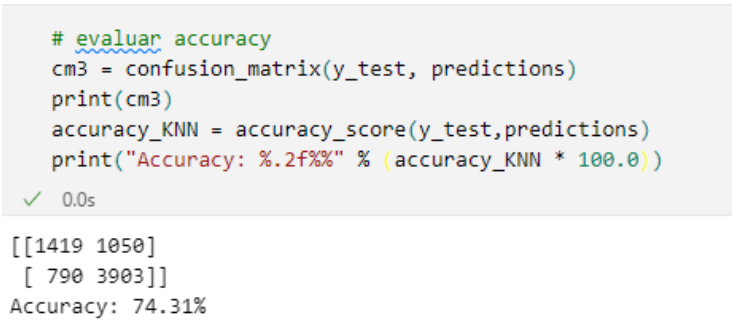
4.5.4.1. Matriz de confusión

La evaluación de las predicciones con K-Nearest Neighbors se realizó a través de la matriz de confusión, como en el caso del modelo anterior.

La ilustración 86 muestra la matriz de confusión.

Análisis:

- La diagonal principal muestra que los verdaderos positivos predichos corresponde a 3.903 casos, mientras que 790 casos se han predicho incorrectamente como positivos, siendo negativos.
- Los verdaderos negativos corresponden a 1.419 casos que se han predicho correctamente, mientras que 1.050 casos se han predicho incorrectamente como negativos, siendo de clase positiva.
- Se ha calculado la métrica accuracy para este modelo con un 74,31% de precisión en las predicciones.



```
# evaluar accuracy
cm3 = confusion_matrix(y_test, predictions)
print(cm3)
accuracy_KNN = accuracy_score(y_test,predictions)
print("Accuracy: %.2f%%" % (accuracy_KNN * 100.0))
```

✓ 0.0s

```
[[1419 1050]
 [ 790 3903]]
Accuracy: 74.31%
```

Ilustración 86 Matriz de confusión de K-Nearest Neighbors, adaptado por (Melo 2024)

4.5.4.2. Otras métricas de evaluación

La ilustración 87 muestra la tabla de métricas de evaluación para el modelo de random forest.

Análisis:

- **Precision:** en este caso nos interesa evaluar la precisión ya que se usa para medir la exactitud de la clase 1 (TotalDeser = 1). En este modelo la precisión es 79% para la clase 1.
- **Recall:** el recall es del 83%. Esto significa que el 83% de todos los ejemplos que realmente pertenecen a la clase 1 fueron correctamente identificados por el modelo.
- **F1-score:** es el promedio armónico de precision y accuracy, donde una puntuación F1 alcanza su mejor valor en 1 (precisión y recuperación perfectas) y su peor valor en 0. El F1-score para este modelo es 81% para la clase (TotalDeser = 1), indicando un buen balance entre precisión y recall para la clase 1.
- **Support:** para la clase 1 hay 4693 ejemplos de la clase 1 en el conjunto de prueba.
- **Accuracy:** el accuracy general del modelo es del 74%, lo que significa que el 74% de todas las predicciones (tanto clase 0 como clase 1) fueron correctas.

```
Reporte de Clasificación de KNN:
              precision    recall  f1-score   support

           0       0.64      0.57      0.61      2469
           1       0.79      0.83      0.81      4693

    accuracy                           0.74      7162
   macro avg       0.72      0.70      0.71      7162
weighted avg       0.74      0.74      0.74      7162
```

Ilustración 87 Otras métricas de evaluación para K-Nearest Neightbors, adaptado por (Melo 2024)

4.5.5. Evaluación de SVM

4.5.5.1. Matriz de confusión

La evaluación de las predicciones con SVM se realizó a través de la matriz de confusión, como en el caso del modelo anterior.

La ilustración 88 muestra la matriz de confusión.

Análisis:

- La diagonal principal muestra que los verdaderos positivos predichos corresponde a 3.865 casos, mientras que 828 casos se han predicho incorrectamente como positivos, siendo negativos.
- Los verdaderos negativos corresponden a 1.509 casos que se han predicho correctamente, mientras que 960 casos se han predicho incorrectamente como negativos, siendo de clase positiva.
- Se ha calculado la métrica accuracy para este modelo con un 67,37% de precisión en las predicciones.

```
# evaluate predictions
cm4 = confusion_matrix(y_test, y_hat_SVM)
print(cm4)

accuracy_SVM = accuracy_score(y_test,y_hat_SVM)
print("Accuracy: %.2f%%" % (accuracy_SVM * 100.0))
```

✓ 0.0s

```
[[ 960 1509]
 [ 828 3865]]
Accuracy: 67.37%
```

Ilustración 88 Matriz de confusión de SVM, adaptado por (Melo 2024)

4.5.5.2. Otras métricas de evaluación

La ilustración 89 muestra la tabla de métricas de evaluación para el modelo de random forest.

Análisis:

- **Precision:** en este caso nos interesa evaluar la precisión ya que se usa para medir la exactitud de la clase 1 (TotalDeser = 1). En este modelo la precisión es 72% para la clase 1.
- **Recall:** el recall es del 82%. Esto significa que el 82% de todos los ejemplos que realmente pertenecen a la clase 1 fueron correctamente identificados por el modelo.
- **F1-score:** es el promedio armónico de precision y accuracy, donde una puntuación F1 alcanza su mejor valor en 1 (precisión y recuperación perfectas) y su peor valor en 0. El F1-score para este modelo es 77% para la clase (TotalDeser = 1), indicando un buen balance entre precisión y recall para la clase 1.
- **Support:** para la clase 1 hay 4693 ejemplos de la clase 1 en el conjunto de prueba.
- **Accuracy:** el accuracy general del modelo es del 67%, lo que significa que el 67% de todas las predicciones (tanto clase 0 como clase 1) fueron correctas.

Reporte de Clasificación de SVM:

	precision	recall	f1-score	support
0	0.54	0.39	0.45	2469
1	0.72	0.82	0.77	4693
accuracy			0.67	7162
macro avg	0.63	0.61	0.61	7162
weighted avg	0.66	0.67	0.66	7162

Ilustración 89 Otras métricas de evaluación para SVM, adaptado por (Melo 2024)

4.5.5.3. Evaluación de MSE y RMSE de los modelos

El MSE o error cuadrático medio y el RMSE o raíz del error cuadrático medio se usan para medir el rendimiento de los modelos de regresión. Mientras que el RMSE es útil para comparar la precisión de diferentes modelos entre sí. Un RMSE más bajo indica que el modelo tiene un mejor ajuste y predice más cerca de los valores reales.

La ilustración 90 muestra los resultados de la evaluación del MSE y RMSE de los modelos presentados.

Análisis:

- **MSE:** en este caso, el MSE más bajo de todos los modelos corresponde a random forest, con MSE de 0.14 significa que, en promedio, el error cuadrado entre las predicciones y los valores reales es pequeño, lo que indica que random forest está haciendo predicciones bastante precisas.
- **RMSE:** un RMSE de 0.38 indica que el error promedio en las predicciones del modelo es de aproximadamente 0.38 unidades (estudiantes). Esto significa que, en promedio, las predicciones del modelo están desviadas del valor real en 0.38 unidades (estudiantes).

```
Decision Tree - MSE: 0.22, RMSE: 0.47
Random Forest - MSE: 0.14, RMSE: 0.38
KNN - MSE: 0.26, RMSE: 0.51
SVM - MSE: 0.33, RMSE: 0.57
```

Ilustración 90 MSE y RMSE de los modelos, adaptado por (Melo 2024)

4.5.6. Síntesis de resultados de los modelos

La tabla 5 muestra el rendimiento de los modelos presentados para una comparación más visual.

Algoritmo	Porcentaje
Regresión logística	0.673694
Decision Tree	0.781346
Random Forest	**0.813181**
K-Nearest Neighbors	0.743089
Support Vector Machine	0.673694

Tabla 5 Síntesis de los resultados de los modelos, adaptado por (Melo 2024)

La ilustración 91 muestra un gráfico sobre el rendimiento de los algoritmos.

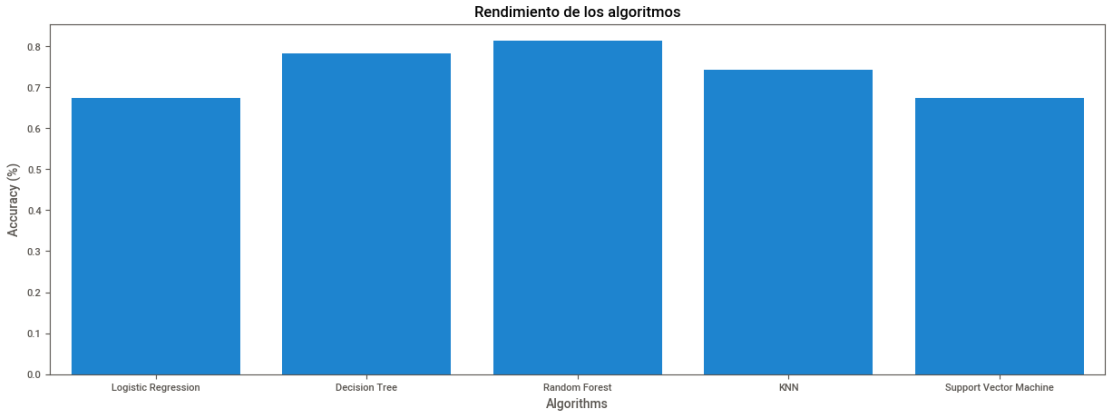


Ilustración 91 Rendimiento de los algoritmos, adaptado por (Melo 2024)

Finalmente, tras el despliegue y evaluación de los modelos presentados (regresión logística, decision tree, random forest, K-Nearest Neighbors y SVM) se concluye que el algoritmo random forest es el mejor modelo para predecir la deserción escolar a nivel de bachillerato, ya que la precisión es del 81,31% en predecir los estudiantes con probabilidad de desertar sus estudios con las variables seleccionadas para la modelización.

4.5.7. Implicaciones prácticas

Este trabajo de investigación es de suma importancia para el sector educativo, pues el conocer el riesgo de deserción en las escuelas puede ser una ventaja para que los tomadores de decisiones redireccionen sus estrategias con el fin de crear y/o mejorar mecanismos de

retención estudiantil en bachillerato, ya que como se observó con la evidencia, las tasas de deserción escolar en el Ecuador se concentran en el nivel de bachillerato.

La deserción escolar es un grave problema socioeconómico que al ser abordado de manera óptima puede beneficiar la vida de muchos estudiantes en el Ecuador, y es crucial que estas estrategias incluyan a toda la comunidad educativa, docentes, directores escolares, padres y madres de familia, para así generar conciencia de la importancia de la formación académica en los adolescentes y jóvenes para asegurar un futuro y calidad de vida digna.

4.5.8. Limitaciones del estudio

Como se ha revisado en la literatura, la mayor parte de algoritmos de clasificación para predecir la deserción escolar se ha realizado mediante el análisis de características sociales, económicas, psicológicas, académicas del estudiante en cuestión. Esta investigación se basó en características a nivel de instituciones educativas, mismas que se encontraban disponibles y abiertas a todo el público en el portal de datos abiertos del MINEDUC. Este particular puede ser un limitante al momento de predecir la deserción escolar, ya que se trabajó con las variables disponibles. No se pudo conseguir información a nivel de estudiante, ya que por temas inherentes de privacidad y protección de la información, el MINEDUC informó a través de una comunicación que el Artículo 21 de la Ley de Estadística indica que:

> "*Los datos individuales que se obtengan para efecto de estadística y censos son de carácter reservado; en consecuencia, no podrán darse a conocer informaciones individuales de ninguna especie, ni podrán ser utilizados para otros fines como de tributación o conscripción, investigaciones judiciales y, en general, para cualquier objeto distinto del propiamente estadístico o censal. Solo se darán a conocer los resúmenes numéricos, las concentraciones globales, las totalizaciones y, en general, los datos impersonales".*

5. CAPÍTULO V

5.1. Conclusiones

El presente trabajo aplicó modelos predictivos de minería de datos para encontrar los factores de deserción escolar. Se utilizaron los siguientes modelos: regresión logística, árboles de decisión, random forest, K-Nearest Neighbors, Support Vector Machine (SVM) con la herramienta Visual Studio Code y Python. Los resultados obtenidos son relevantes para la toma decisiones de cómo mejorar los procesos de enseñanza y aprendizaje del sistema educativo nacional, y por ende mejorar las condiciones educativas de los niños, niñas y adolescentes.

El presente estudio evidencia los factores que pueden ser causa de la deserción escolar de los estudiantes de bachillerato a nivel nacional, siendo los principales factores que explican la deserción escolar, el número de docentes de género masculino, el número de estudiantes de

bachillerato de género masculino, el número de estudiantes promovidos de género masculino, el número de estudiantes no promovidos de género masculino, jornadas matutina y vespertina y la jornada matutina, vespertina y nocturna. Todas estas variables mostraron una correlación positiva con la deserción escolar. Los resultados muestran que los estudiantes tienden a desertar cuanto existen más docentes de género masculino, cuando el estudiante es de género masculino, o cuando el estudiante es o no promovido de año escolar. Las variables de jornada escolar mostraron que estudiantes que asistan a la escuela en estas dos jornadas que ofrece la escuela tienen más probabilidad de desertar sus estudios que estudiantes que asistan a otras jornadas. La variable sostenimiento de la escuela particular mostró una correlación negativa, lo que significa que a medida que el tipo de sostenimiento particular en las escuelas aumente, la deserción escolar disminuirá.

La selección de variables fue relevante como preparación a la etapa de minería de datos y se utilizaron estrategias para evitar la multicolinealidad y por ende, que las variables predictoras no tengan alta correlación entre ellas. Este proceso ayudó a que los modelos presenten predicciones certeras y reales, sin sobre ajuste o bajo ajuste. En este caso, las variables número de docentes de género femenino, el número de estudiantes de bachillerato de género femenino, el número de estudiantes promovidos de género femenino, el número de estudiantes no promovidos de género femenino tenían una alta correlación con la variable de respuesta total de estudiantes desertores, sin embargo estaban altamente correlacionadas con otras variables predictoras, lo que ocasionaba multicolinealidad, por lo que no se consideraron para el modelado de los datos.

Se aplicó exitosamente varios algoritmos de aprendizaje automático supervisado los cuales ayudaron a predecir los factores de deserción escolar de manera óptima, y su interpretación y evaluación fue crucial para determinar el mejor modelo en predecir la deserción, el cual fue random forest, con una precisión del 81% en sus predicciones con un error cuadrático medio de 0.14.

Se concluye que la etapa más importante para el modelado de algoritmos de clasificación es la etapa de preprocesamiento y transformación. Los datos proporcionados por el MINEDUC traían varios errores de codificación de las etiquetas y nombres de variables, por lo que se tuvo que hacer un análisis minucioso de cada variable contenida en los archivos para luego pasar al preprocesamiento de los datos. la metodología KDD y su enfoque fue determinante para realizar este proceso, ya que se aplicó técnicas para remover datos ruidosos, seleccionar estrategias para manejo de valores perdidos o nulos entre otras.

5.2. Recomendaciones

Se recomienda continuar analizando los factores de deserción escolar en Ecuador mediante técnicas de minería de datos y de la mano de expertos educativos, con mejor y más información para dar una respuesta integral, que beneficie a todos los niños, niñas y adolescentes del país. Además, enfocarse en los factores de deserción escolar de estudiantes de bachillerato es muy importante pues es la etapa que posteriormente les permitirá continuar sus estudios en la educación superior. Esta transición de bachillerato a educación superior es importante pues prepara a los estudiantes con las competencias necesarias para especializarse en la educación superior.

Se recomienda la implementación y actualización constante de los sistemas de alerta temprana para prevenir la deserción escolar en escuelas donde se conoce que son más propensas a que sus estudiantes abandonen los estudios. Ecuador cuenta con un sistema de alerta temprana que puede ser de utilidad para intervenir en casos donde se puede dar deserción.

Se recomienda ampliamente continuar con los programas de búsqueda activa escolar que implementa el MINEDUC en alianza con otros actores del sector educativo, tales como "Todos al Aula" pues fomenta la reinserción escolar y promueve el acceso al sistema educativo formal de los niños, niñas y adolescentes que se encuentran fuera del sistema educativo. A pesar de que no es una estrategia propia para frenar la deserción, la inclusión de los estudiantes en el sistema formal puede promover motivación a que otros estudiantes no abandonen los estudios. Existen experiencias valiosas de búsqueda activa escolar de las que el país podría beneficiarse y así disminuir la deserción escolar.

6. ANEXOS

Anexo 1 Cronograma de actividades

	JUNIO			JULIO		
Actividad	**10-14**	**17-21**	**24-28**	**1-5**	**8-12**	**15**
1. Preprocesamiento de datos						
2. Análisis exploratorio de datos						
3. Desarrollo, evaluación y refinamiento del modelo predictivo						
4. Visualización de resultados						
5. Desarrollo de segundo modelo						
6. Visualización de los datos						
7. Elaboración del documento escrito						

Anexo 2 cuaderno de Visual Studio Code

- Enlace al caderno

7. REFERENCIAS BIBLIOGRÁFICAS

Alban Taipe, M. S., & Sánchez, D. M. (2018). Prediction of university dropout through technological factors: a case study in Ecuador. *Revista Espacios*, 1-7. Obtenido de https://www.revistaespacios.com/a18v39n52/a18v39n52p08.pdf

Comisión Económica para América Latina y el Caribe CEPAL. (Agosto de 2020). *La educación en tiempos de la pandemia de COVID-19.* Obtenido de https://repositorio.cepal.org/server/api/core/bitstreams/c29b3843-bd8f-4796-8c6d-5fcb9c139449/content

Cornejo Sifuente, S. G., Naranjo Cantabrana, M. G., Ávila Santana, F. A., Vega Pérez, L. G., Osúa Acosta, I. F., & Sotomayor Fierro, M. d. (Septiembre-Octubre de 2023). Modelo Predictivo de la Deserción Escolar en Educación Superior: una Aproximación desde la Minería de Datos Utilizando la Metodología CRISP-DM. *Ciencia Latina Revista Científica Multidisciplinar*, 1-16. doi:10.37811/cl_rcm.v7i5.8363

Cuji, B., Gavilanes, W., & Sánchez, R. (Julio de 2017). Modelo predictivo de deserción estudiantil basado en árboles de decisión. *Revista Espacios*, 1-9. Obtenido de Modelo predictivo de deserción estudiantil basado en árboles de decisión: https://www.revistaespacios.com/a17v38n55/a17v38n55p17.pdf

Fondo de las Naciones Unidas para la Infancia, UNICEF. (8 de Febrero de 2021). *Informes sobre la situación de la educación durante COVID-19.* Obtenido de LACRO COVID-19 Respuesta educativa: UPDATE 21 Estado de la reapertura de escuelas: https://www.unicef.org/lac/media/20541/file

Forbes Ecuador. (16 de Septiembre de 2022). *Forbes EC*. Obtenido de Deserción escolar: una realidad que destruye el futuro: https://www.forbes.com.ec/columnistas/desercion-escolar-una-realidad-destruye-futuro-n22171

Garača, Ž., & Čukušić, M. (Enero de 2010). *Student Dropout Analysis with Application of Data Mining Methods.* Obtenido de https://www.researchgate.net/publication/44239145_Student_Dropout_Analysis_with_Application_of_Data_Mining_Methods

Han, J., Kamber, M., & Pei, J. (2001). *Data mining concepts and techniques.* San Francisco: Elsevier Inc. doi:10.4236/jcc.2014.214002

Hernández González, A. G., Melendez Armenta, R., Morales Rosales, L. A., García Barrientos, A., Tecpanecatl Xihuitl, J., & Algredo-Badillo, I. (Noviembre de 2016). Comparative Study of Algorithms to Predict the Desertion in the Students at the ITSM-Mexico. *IEEE Latin America Transactions*. doi:10.1109/TLA.2016.7795831

Instituto Nacional de Evaluación Educativa. (2018). *La educación en Ecuador: logros alcanzados y nuevos desafíos Resultados educativos 2017-2018.* Quito. Obtenido de https://evaluaciones.evaluacion.gob.ec/BI/la-educacion-en-ecuador-logros-alcanzados-y-nuevos-desafios-resultados-educativos-2017-2018/

Kumar, M., Singh, A. J., & Handa, D. (Marzo de 2017). *Modern Education and Computer Science Press.* doi:10.5815/ijeme.2017.02.02

Llaugel, F. A., & González-Disla, R. R. (Febrero de 2016). Un Modelo Predictivo de Deserción Escolar en la República Dominicana. 1-29. doi:10.13140/RG.2.2.10903.37283

Loaiza, D., Romero, J., Ronquillo, P., García, I., & Diaz, M. (Marzo-Abril de 2023). Identificación de los factores de la deserción académica en el sistema educativo del Ecuador. *Ciencia Latina Revista Científica Multidisciplinar,* 2-16. doi:10.37811/cl_rcm.v7i2.6190

Manandhar, N., & Sthapit, A. B. (Enero de 2005). Logistic Regression Model For Primary School Dropout Children Of Chitwan District Of Nepal. *RepEc*. Obtenido de Logistic Regression Model For Primary School Dropout Children Of Chitwan District Of Nepal: https://www.researchgate.net/publication/227368344_Logistic_Regression_Model_For_Primary_School_Dropout_Children_Of_Chitwan_District_Of_Nepal?enrichId=rgreq-92901d5b1b7690155a514c5654ee4cfb-XXX&enrichSource=Y292ZXJQYWdlOzIyNzM2ODM0NDtBUzo3MDIxODIwNzI1NDUy

Márquez, C., Ventura, S., & Romero, C. (2011). Predicting School Failure Using Data Mining. 1-5. Obtenido de https://www.researchgate.net/publication/221570432_Predicting_School_Failure_Using_Data_Mining

Merchan, S., & Duarte García, J. A. (Junio de 2016). *Research Gate.* doi:10.1109/TLA.2016.7555255

Ministerio de Educación de Ecuador. (2015). *Ley Orgánica de Educación Intercultural.* Quito. Obtenido de https://educacion.gob.ec/wp-content/uploads/downloads/2017/02/Ley_Organica_de_Educacion_Intercultural_LOEI_codificado.pdf

Ministerio de Educación de Ecuador. (2020). *Informe Preliminar.* Obtenido de [Rendición de Cuentas]: https://educacion.gob.ec/wp-content/uploads/downloads/2021/05/Informe-preliminar-RC-2020.pdf

Ministerio de Educación de Ecuador. (Noviembre de 2021). *Datos Abiertos Estadística Educativa Volumen 4.* Obtenido de https://educacion.gob.ec/datos-abiertos/

Ministerio de Educación de Ecuador. (Octubre de 2021). *Ficha metodológica del indicador tasa de abandono escolar.* Obtenido de educacion.gob.ec: https://educacion.gob.ec/wp-content/uploads/downloads/2021/10/Abandono-Escolar.pdf

Ministerio de Educación de Ecuador MINEDUC. (2 de Junio de 2021). *Se presentó los cinco ejes de trabajo del Ministerio de Educación*. Obtenido de educacion.gob.ec: https://educacion.gob.ec/se-presento-los-cinco-ejes-de-trabajo-del-ministerio-de-educacion/

Nigro, H. O., Xodo, D., Corti, G., & Terren, D. (s.f.). *KDD (Knowledge Discovery in Databases): Un proceso centrado en el usuario.* Obtenido de http://sedici.unlp.edu.ar/bitstream/handle/10915/21220/Documento_completo.pdf?sequence=1

Pachay, M. J., & Rodríguez, M. (4 de Enero de 2021). La deserción escolar: Una perspectiva compleja en tiempos de pandemia. *Polo del conocimiento Opern journal systems*. Obtenido de La deserción escolar: Una perspectiva compleja en tiempos de pandemia: https://polodelconocimiento.com/ojs/index.php/es/article/view/2129/4240

Pérez, C., & Santín, D. (2008). *Minería de datos Técnicas y herramientas.* Madrid: Closas Orcoyen, S.L.

Quinlan, J. (1986). Induction of decision trees. *Machine Learning*. doi:https://doi.org/10.1007/BF00116251

Román, M. (2013). *Dialnet.* Recuperado el 15 de Diciembre de 2023, de Factores asociados al abandono y la deserción escolar en América Latina: una mirada en conjunto: https://dialnet.unirioja.es/servlet/articulo?codigo=4453200

Solís Ventura, J. C., Quiroz Fernández, S., & Fosado Téllez, O. (Septiembre-Diciembre de 2022). Modelo de regresión logística para la estimación de la deserción escolar del posgrado en la Universidad Técnica de Manabí, Ecuador. *Revista Bases de la Ciencia*, 1-14. doi:https://doi.org/10.33936/revbasdelaciencia.v7i3.5197

Sotomonte Castro, J. E., Rodríguez Rodríguez, C. C., Montenegro Marín, C. E., Gaona García, P. A., & Castellanos, J. G. (Octubre de 2016). Hacia la construcción de un modelo predictivo de deserción académica basado en técnicas de minería de datos. *Revista Científica*, 1-15. doi:10.14483/23448350.11089

Statsmodels Variance inflation factor. (14 de Diciembre de 2023). Obtenido de https://www.statsmodels.org/stable/generated/statsmodels.stats.outliers_influence.variance_inflation_factor.html

Theodoridis, S., & Koutroumbas, k. (Marzo de 2006). Pattern Recognition. *IEEE Transactions on Neural Networks*. Obtenido de Pattern Recognition: https://www.researchgate.net/publication/3304050_Pattern_Recognition_Theodoridis_S_and_Koutroumbas_K_2006

Timarán-Pereira, S. R., Hernández-Arteaga, I., Caicedo-Zambrano, S. J., Hidalgo-Troya, A., & Alvarado-Pérez, J. C. (2016). *El proceso de descubrimiento de conocimiento en bases de datos. En Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional.* Bogotá: Ediciones Universidad Cooperativa de Colombia. doi:http://dx.doi.org /10.16925/9789587600490

United Nations. (10 de Diciembre de 1948). *Universal Declaration of Human Rights*. Obtenido de https://www.un.org/en/about-us/universal-declaration-of-human-rights

Zambrano Ostaiza, C. J., & Guaña Bravo, P. V. (Marzo de 2024). La iniciación Laboral Temprana y la Deserción Escolar en los Estudiantes de Bachillerato. *Ciencia Latina Revista Científica Multidisciplinar*, 4-12. doi:10.37811/cl_rcm.v8i1.10121