ESCUELA DE NEGOCIOS

MAESTRÍA EN INTELIGENCIA DE NEGOCIOS Y CIENCIA DE DATOS

DESERCIÓN ESCOLAR EN ECUADOR 2022. MODELO BASADO EN ÁRBOLES DE DECISIÓN

Profesor
Víctor Manuel González Holguín

Autor
Carol Solange Morales Gavilanes

2024

Resumen

A nivel nacional, según cifras del INEC para el 2022, un total de 4,309,139 estudiantes estuvieron matriculados durante el periodo 2021-2022. De este grupo, 90,850 estudiantes decidieron abandonar sus estudios, lo que representa una tasa de deserción escolar del 2.11 %, la más alta registrada en los últimos cuatro periodos escolares. Estos datos evidencian un aumento en el número de estudiantes que optaron por interrumpir su educación, con repercusiones tanto a nivel individual, como problemas de salud mental, menor calidad de vida y limitaciones en el desarrollo personal, así como a nivel social, con costos económicos en términos de pérdida de productividad, mayores tasas de desempleo y un aumento en la propensión a la delincuencia, lo que impacta negativamente en la seguridad pública y la cohesión social. En este contexto, la presente investigación tiene como objetivo clasificar a los estudiantes con mayor riesgo de abandonar la escuela, considerando factores demográficos y socioeconómicos. Para lograr esto, se empleó el árbol de decisión, una técnica de modelado predictivo en el campo del aprendizaje automático que permite clasificar instancias o predecir valores basándose en la observación de atributos. Los resultados obtenidos revelaron un modelo con una precisión del 85 %, destacando la importancia de características como la edad, el estado civil, el sexo y el ingreso per cápita en la identificación de patrones y relaciones que contribuyeron a la clasificación de la deserción escolar entre los estudiantes en Ecuador.

Palabras claves: Deserción escolar, árbol de decisión, clasificación.

Abstract

At the national level, according to INEC figures for 2022, a total of 4,309,139 students were enrolled during the 2021-2022 period. From this group, 90,850 students decided to drop out of their studies, representing a dropout rate of 2.11 %, the highest recorded in the last four school periods. These data highlight an increase in the number of students who chose to interrupt their education, with repercussions both at the individual level, such as mental health problems, lower quality of life, and limitations in personal development, as well as at the social level, with economic costs in terms of loss of productivity, higher unemployment rates, and an increase in the propensity for crime, negatively impacting public safety and social cohesion. In this context, the present research aims to classify students at higher risk of dropping out of school, considering demographic and socioeconomic factors. To achieve this, the decision tree technique was employed, a predictive modeling technique in the field of machine learning that allows for the classification of instances or prediction of values based on attribute observation. The results obtained revealed a model with an accuracy of 85 %, highlighting the importance of characteristics such as age, marital status, gender, and per capita income in identifying patterns and relationships that contributed to the classification of school dropout among students in Ecuador.

Keywords: School dropout, decision tree, classification.

Índice General

Índice de Figuras

Índice de Tablas

Capítulo 1

Introducción

El reconocimiento del derecho global a la educación, tal como lo establece el Artículo 26 de la Declaración Universal de Derechos Humanos, constituye el núcleo fundamental sobre la necesidad de garantizar a todos un acceso pleno e igualitario a la educación. Este compromiso tiene como objetivo fomentar el ideal de igualdad de oportunidades en el ámbito educativo (UNESCO, 2011).

En el Ecuador, la Constitución del Ecuador (2008) en el Artículo 28 establece que la educación debe servir al interés público y no a intereses individuales. Se garantiza el acceso universal, permanencia y movilidad sin discriminación en los niveles inicial, básico y bachillerato o equivalentes. La educación, obligatoria en los niveles mencionados, se considera un derecho de toda persona para interactuar entre culturas y participar en una sociedad que aprende.

A pesar de ello, un problema que sigue preocupando a la sociedad es la salida prematura de niños y adolescentes del sistema educativo. Según datos de la Comisión Económica para América Latina y el Caribe (CEPAL), en la región se registra un promedio del 10 % de

abandono educativo entre adolescentes de 12 a 17 años, Sin embargo, al dirigir la atención hacia el grupo de jóvenes de 18 a 24 años, las tasas de abandono se elevan considerablemente, alcanzando un promedio regional del 24 %. Este aumento sugiere una transición crítica en la retención educativa durante los años posteriores a la adolescencia (Rossel et al., 2022).

En este sentido, el estudio se enfoca en analizar la relación entre factores sociodemográficos y la decisión de deserción o permanencia en el sistema educativo. La deserción escolar, un desafío significativo con implicaciones tanto individuales como sociales, motiva la necesidad de comprender en detalle los elementos que la influyen. Para ello, se seleccionaron factores socioeconómicos y demográficos como variables clave basándose en la evidencia de la literatura académica. La investigación busca identificar patrones y relaciones entre estas variables para prever y clasificar la deserción escolar.

Asi, el estudio se compone de cinco capítulos. En el primer capítulo, se aborda la problemática de la deserción escolar en Ecuador, proporcionando la motivación del estudio y delineando los objetivos de investigación.

En el segundo capítulo, se lleva a cabo una conceptualización general de la deserción escolar, seguido por una revisión empírica de estudios previos sobre el abandono escolar. Este capítulo concluye con un análisis de la situación actual del abandono escolar y su evolución a lo largo del tiempo.

El tercer capítulo detalla la descripción de la base de datos derivada de la Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU) del año 2022. Se presenta la metodología utilizada y se describen las variables empleadas en el modelo de estudio.

En el cuarto capítulo, se exponen los resultados obtenidos mediante el modelo propuesto. Finalmente, en el quinto capítulo, se exponen las conclusiones derivadas de la investigación, acompañadas de posibles líneas de investigación para el futuro.

1.1. Planteamiento del Problema

El fenómeno de la deserción escolar representa un desafío significativo en el ámbito educativo, con repercusiones a nivel individual, social y económico. La decisión de un estudiante de abandonar prematuramente sus estudios puede tener consecuencias a nivel personal, incluyendo un impacto negativo en la autoestima, la salud mental y las perspectivas laborales. A nivel económico, la deserción escolar conlleva costos significativos en aspectos fiscales relacionados con el gasto en servicios sociales, prevención de la criminalidad, formación laboral y se reduce el potencial de contribución al desarrollo económico (Espínola y Claro, 2010; Levin, 2003).

En Ecuador, la culminación exitosa de la educación conlleva beneficios significativos para los individuos, evidenciados tanto en su productividad laboral como en sus ingresos. A medida que avanzan en su ciclo laboral, aquellos con mayores niveles de capital humano y educación experimentan aumentos salariales más notables (Tarupi, 2015). Según datos del Instituto Nacional de Estadística y Censos (INEC, 2021), la relación entre nivel educativo y empleo adecuado es evidente, destacando que el 56,2 % de las personas con educación superior tienen empleo adecuado, en comparación con el 8,6 % de aquellos sin instrucción formal.

Asimismo, los datos disponibles revelan que, en los quintiles más altos, los años promedio de escolaridad son significativamente mayores. Por ejemplo, en el quintil 5, la media de años de educación es de 13,5, mientras que en el quintil 1 es de solo 8,2 años. Esta disparidad subraya la existencia de desafíos estructurales que pueden perpetuar las inequidades socioeconómicas (INEC, 2021).

En lo que respecta a los costos sociales, la creciente disparidad en las remuneraciones entre trabajadores con educación universitaria y aquellos sin ella ha tenido un impacto

adverso en la distribución del ingreso a largo plazo, lo cual incide directamente en los problemas de cohesión social, así como en las dificultades relacionadas con la gobernabilidad. Un aspecto clave de este desafío es el costo social asociado a la baja escolaridad, que se manifiesta como un factor significativo en la inequidad económica y social, afectando negativamente la estabilidad y armonía en la sociedad (Ocampo y CEPAL, 2021).

En el periodo lectivo 2021-2022, el sistema educativo ecuatoriano experimentó una deserción escolar del 2.11 %, lo que equivale a 90,850 de 4'309.139 estudiantes que no completaron el año académico, según datos proporcionados por el Ministerio de Educación. Este fenómeno conlleva a una reducción del 3 % en la tasa de matriculación [1] estudiantil a nivel nacional. Entre los factores asociados a la deserción, se identifican diversas causas, destacándose la insuficiencia de recursos económicos con un porcentaje significativo del 24.5 %. Asimismo, se evidencia una falta de interés por parte de los estudiantes, representando un 23 % de incidencia en la problemática. La condición de padres adolescentes, con un 9 %, también figura como un factor influyente en la descontinuación de la educación formal (MINEDUC, 2022). Estos hallazgos resaltan la complejidad y la multi-causalidad de la deserción escolar en el contexto ecuatoriano.

[1] La tasa de matrícula refleja la proporción de estudiantes dentro del rango de edad apropiado para cursar un determinado nivel educativo que están matriculados en ese mismo nivel (MINEDUC, 2022).

1.2. Identificación del Objeto de Estudio

El trabajo de investigación se centra en analizar la relación entre factores socioeconómicos y demográficos que influyen en la decisión de deserción o permanencia en el sistema educativo. La deserción estudiantil representa un desafío significativo que afecta no solo a los individuos involucrados, sino también a la sociedad en su conjunto. La comprensión detallada de los elementos que inciden en esta problemática es esencial para desarrollar estrategias eficaces que contribuyan a la retención y culminación exitosa de la educación (Lugo, 2013).

La selección de los factores socioeconómicos y demográficos como variables de interés se sustenta en la evidencia de la literatura académica que sugiere su influencia en la deserción escolar. Estos factores pueden abordar cuestiones como el nivel socioeconómico familiar, la ubicación geográfica y otros aspectos relacionados. La inclusión de estas variables posibilitará identificar patrones y relaciones que permitan predecir y clasificar la deserción escolar.

En este estudio, la variable de interés será binaria, dividiendo a los individuos en dos grupos distintos: los que abandonan sus estudios y los que no lo hacen, proporcionando así una base clara para la aplicación de un modelo supervisado en el que el algoritmo se entrena utilizando un conjunto de datos etiquetados (Nilsson, 2005).

Esta investigación tiene como objetivo contribuir desde una perspectiva descriptiva y mediante la aplicación de modelos supervisados de machine learning, a identificar las características individuales y contextuales más relevantes que influyen en la deserción estudiantil en la educación en general de los estudiantes ecuatorianos. Utilizando técnicas avanzadas de aprendizaje automático, se pretende analizar y predecir patrones de deserción, incorporando variables específicas que puedan influir en las decisiones de los estudiantes, de manera que describan los perfiles de los estudiantes propensos a desertar.

1.3. Objetivo general

Identificar patrones y factores que influyen en la deserción escolar, clasificando a los individuos según su propensión a abandonar el sistema educativo mediante técnicas de minería de datos supervisadas aplicadas a variables socioeconómicas y demográficas.

1.4. Objetivos específicos

- Seleccionar y recopilar un conjunto de datos representativo que incluya variables socioeconómicas y demográficas relevantes, influyentes en la deserción escolar.
- Identificar posibles patrones preliminares para entender la distribución y relaciones entre las variables seleccionadas mediante un análisis exploratorio de datos.
- Asegurar la calidad y consistencia del conjunto de datos, aplicando técnicas preprocesamiento de datos para abordar problemas como datos faltantes o atípicos.
- Construir un modelo de árbol de decisión para predecir la deserción escolar, implementando técnicas de minería de datos supervisadas.

Capítulo 2

Revisión de la Literatura

Según la teoría del Capital Humano de Becker (1962), se introduce al capital humano como una forma de inversión que aumenta la productividad y los ingresos individuales a lo largo del tiempo. Este enfoque implica que la toma de decisiones educativas se rige por un cálculo racional, donde los individuos evalúan los costos y beneficios de invertir en su formación. La inversión en educación se presenta como el camino estratégico para potenciar las habilidades de la población, considerando que las personas son el recurso principal y el factor pensante de la sociedad en general (Coronel, 2010).

En este sentido, la deserción estudiantil, conceptualizada de diversas maneras por académicos y expertos, se percibe como un fenómeno intrínseco a los sistemas educativos que va más allá de la mera interrupción de estudios. Desde la perspectiva de varios autores, como Salazar (2017), Silvera (2016) la deserción se entiende como el resultado de una compleja interacción entre factores académicos, sociales, personales y política educativa que influyen en la decisión del estudiante de abandonar sus estudios. La deserción estudiantil ha adquirido una relevancia significativa como problema social, dado que su impacto trasciende el ámbito

educativo y afecta directamente al desarrollo de la sociedad (Lugo, 2013; Moreno y Moreno, 2013).

A lo largo del tiempo, la preocupación por la deserción estudiantil ha sido objeto de atención tanto por parte de investigadores como de educadores. Este fenómeno, dada su complejidad, ha motivado la realización de diversos estudios con el objetivo de arrojar luz sobre las causas fundamentales que llevan a la interrupción de la trayectoria educativa de los estudiantes (Castillo et al., 2014). Estas investigaciones han abordado el tema desde diferentes perspectivas, explorando las dimensiones académicas, sociales y personales implicadas en el proceso de deserción.

Rochin (2021) aborda diversas causas relacionadas con la deserción escolar, entre las que se incluyen deficiencias en los programas de estudio, insuficiencias en la preparación y actualización del personal docente, obstáculos familiares enfrentados por los estudiantes, falta de un objetivo o proyecto de vida claro por parte de estos. El autor concluye que la deserción escolar se atribuye a diversos factores, abarcando aspectos personales como la falta de motivación y dificultades en las relaciones interpersonales, así como situaciones específicas como embarazos en la adolescencia. Además, se identifican problemas de índole socioeconómica ligados al nivel de ingreso familiar.

En el estudio desarrollado por Román (2013), se considera que la deserción escolar está vinculada al nivel socioeconómico, siendo la pobreza un factor determinante en este fenómeno, el contexto geográfico, especialmente en entornos rurales, y el capital cultural en el ámbito familiar también influyen significativamente. Menciona además que, la escolaridad de los padres, las expectativas académicas y la estructura familiar, son factores que afectan la conclusión de la educación. Por su parte, Beyer (1998) señala que los ingresos familiares y las expectativas educativas están fuertemente correlacionados y que la deserción tiene sus raíces en la percepción de que la educación no ofrece recompensas suficientes o en la necesidad de

cubrir urgencias económicas familiares, por lo que incorporarse al mercado laborar resulta atractivo.

La deserción escolar tiene diversas causas, incluyendo la distancia a los colegios, la carencia de transporte escolar en áreas rurales, dificultades económicas, especialmente cuando las familias tienen recursos limitados para cubrir los gastos educativos, y problemas académicos. La deserción también está vinculada a las instituciones educativas mismas, especialmente cuando los costos de las matrículas son elevados, estos factores son identificados como barreras significativas que promueven el abandono del sistema educativo (Moreno y Moreno, 2013).

Recientemente, ha surgido un creciente interés en la aplicación de técnicas de minería de datos en el ámbito educativo, enfocándose particularmente en el análisis de aspectos como la deserción escolar. El propósito es rastrear las acciones de los estudiantes, identificando de manera oportuna cambios en su comportamiento académico que puedan prever, una eventual deserción o abandono escolar. En este sentido, Pérez et al. (2018) abordaron el análisis predictivo de la deserción estudiantil en Colombia aplicando técnicas de minería de datos, en este, se determinaron factores demográficos y registros académicos asociados con las tasas de no finalización de estudiantes.

En la misma línea, en el estudio desarrollado por Valero et al. (2005) contemplan técnicas de minería de datos, centrándose en la construcción de un modelo predictivo utilizando el algoritmo de árboles de clasificación para abordar la problemática de la deserción de estudiantes. Comparando el desempeño de este algoritmo con el de los k vecinos más cercanos, se evidenció que el árbol de clasificación logró una mejor precisión y entre los hallazgos se identificaron tres causas principales de deserción estudiantil: la edad, los ingresos familiares y el nivel de inglés.

La investigación de Cerón y Verduzco (2005) respalda la idoneidad de los modelos de regresión, especialmente árboles de decisión y redes neuronales, para abordar problemas predictivos, en el estudio se incluyeron variables como la edad, sexo, comunidad y estado civil, y sus hallazgos revelaron un rendimiento destacado de ambos modelos, evidenciando su capacidad para capturar relaciones no lineales y adaptarse a la variabilidad inherente a los problemas de predicción. Asimismo, Cuji et al. (2017), López et al. (2013) y Ramírez et al. (2022) coinciden con la aplicación de árboles de decisión para desarrollar un modelo predictivo, los autores utilizaron la misma técnica incluyendo variables socioeconómicas y académicas identificando variable como la edad y las calificaciones como las principales influencias en la deserción.

En otras investigaciones se utilizan técnicas como regresiones logísticas para predecir la probabilidad de deserción estudiantil (Gómez et al., 2016; Solís et al., 2022), redes bayesianas como método de clasificación supervisada que introduce relaciones de dependencia entre características mediante la representación de un grafo (Eckert y Suénaga, 2015; García et al., 2022) y redes neuronales (Cerón y Verduzco, 2017).

Calero et al. (2023) utilizando datos de la Encuesta Nacional de Hogares empleó un modelo de regresión logística para la estimación. Los resultados revelaron que la probabilidad de deserción escolar se incrementa significativamente en relación con diversos factores. Desde la perspectiva socioeconómica, se encontró un mayor riesgo de deserción para estudiantes pertenecientes a hogares de nivel socioeconómico bajo, aquellos que viven en familias monoparentales y cuyos padres han desertado de la educación básica. En cuanto a los factores personales, la probabilidad de deserción aumenta con la edad del estudiante, si es de sexo masculino y si tiene la necesidad de trabajar.

La relación negativa entre la deserción escolar y tener padres con mayor nivel educacional, ser mujer, residir en una zona urbana y tener un ingreso per cápita mayor es un resultado frecuente obtenido por diversos estudios en los que se señala que estos factores

disminuyen la probabilidad de deserción (Beyer, 1998; Melis et al., 2005; Román, 2013). Por el contrario, otros estudios encontraron que el sexo de una persona y la zona de residencia no influyen en la probabilidad de la decisión de desertar (Sapelli y Torche, 2004).

Por otro lado, en un estudio realizado para el caso chileno se verifican los factores intraescolares que inciden en la deserción escolar, específicamente en contextos de alta vulnerabilidad con niveles significativos de pobreza, los autores hacen una revisión y en base a la experiencia educativa, se revela que los estudiantes con bajo rendimiento académico, desmotivación, problemas conductuales e inestabilidad escolar pueden ser individuos críticos a la hora de tomar la decisión de abandonar la escuela (Espinoza et al., 2014). Un aspecto que los autores recalcan es la falta de apoyo tanto de la familia como de la institución educativa, por lo que se destaca la necesidad de intervenir no solo a nivel académico, sino emocional y de apoyo social. como resultado del trabajo se sugiere intervenir y aplicar políticas efectivas en sectores sociales con niveles de pobreza similares a los del estudio, con programas de retención y reescolarización.

Para finalizar, la deserción estudiantil considera varios factores como la causa de este problema social, el hecho de que la educación sea vista como una inversión clave para aumentar la productividad y los ingresos individuales a lo largo del tiempo agrava aun más la situación, pues la toma de decisiones educativas, van ligados a los costos y beneficios de invertir en formación. Por lo tanto, la deserción estudiantil se conceptualiza como un fenómeno complejo, influenciado por factores académicos, sociales, personales y de política pública. Las investigaciones recientes aplican técnicas de minería de datos para identificar patrones y factores predictivos de la deserción, destacando variables demográficas y socioeconómicas claves. En particular, se evidencia la necesidad de intervenciones integrales, no solo académicas, sino también emocionales y de apoyo social, especialmente en contextos de alta vulnerabilidad, para abordar eficazmente este desafío educativo.

2.1. Deserción escolar en el Ecuador

El sistema educativo ecuatoriano ha experimentado diversas fases desde 1989, y desde el año 2011, el currículo para el Bachillerato General Unificado ha estado en vigencia mediante el acuerdo Ministerial Nro. 242-11. Actualmente, el sistema educativo se estructura en varios niveles, comenzando con la Educación Inicial, seguida por la Educación General Básica y el Bachillerato. La educación obligatoria abarca estos tres primeros niveles. Posteriormente, se accede a los estudios superiores, donde los estudiantes tienen la libertad de elegir la carrera que deseen, alineándola con sus intereses personales (MINEDUC, 2016).

En la Tabla 2.1, se puede observar la evolución de la tasa de abandono escolar [2] durante el periodo 2009 - 2022. Las cifras muestran que desde el año 2009, la tasa de deserción escolar ha tenido una reducción en 2,6 % hasta el año 2019, sin embargo, a partir de este mismo año, ha empezado a crecer, alcanzando el 2,11 % en 2021.

[2]Corresponde al porcentaje de estudiantes que abandonan un curso específico al término de un periodo escolar. Este cálculo se realiza en relación con el total de estudiantes matriculados al final de ese curso y periodo escolar.

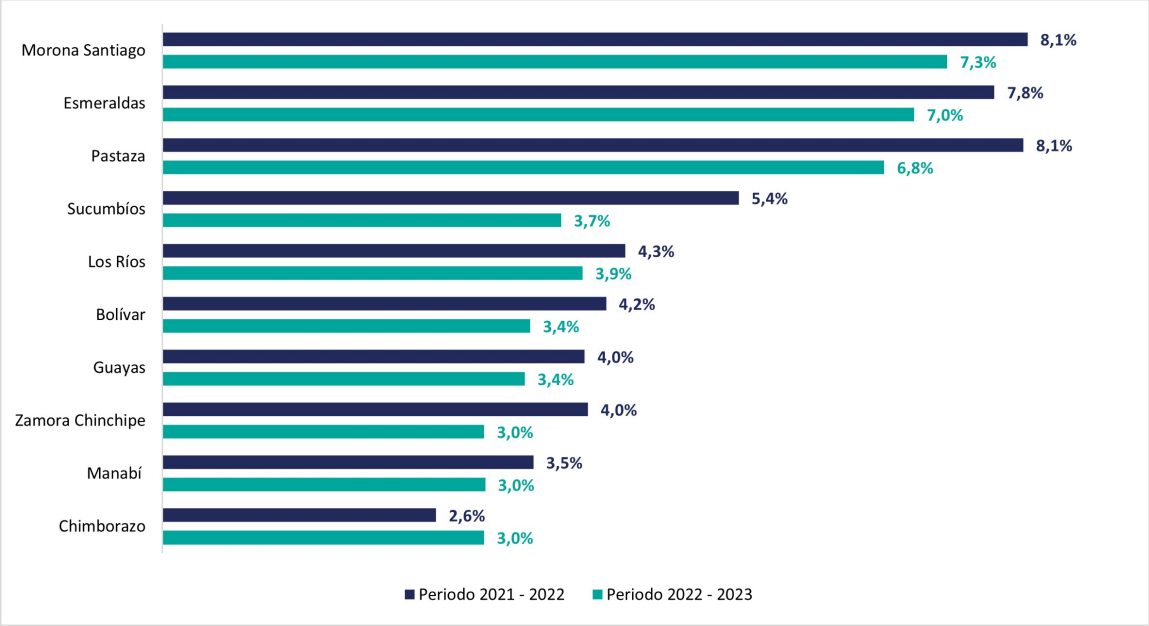
Tabla 2.1: Evolución de la tasa de abandono escolar

Periodo	Desertores	Matriculados	% Tasa de abandono
2009 - 2010	178.022	4.103.224	4,34 %
2010 - 2011	198.994	4.227.768	4,71 %
2011 - 2012	245.552	4.380.546	5,61 %
2012 - 2013	232.089	4.449.216	5,22 %
2013 - 2014	177.588	4.560.138	3,89 %
2014 - 2015	143.343	4.728.582	3,03 %
2015 - 2016	128.338	4.627.946	2,77 %
2016 - 2017	128.674	4.581.883	2,81 %
2017 - 2018	103.681	4.516.067	2,30 %
2018 - 2019	92.457	4.473.021	2,07 %
2019 - 2020	76.338	4.407.030	1,73 %
2020 - 2021	76.426	4.314.777	1,77 %
2021 - 2022	90.850	4.309.139	2,11 %

Fuente: MINEDUC, 2022
Elaboración propia

Por otra parte, en el ciclo 2022-2023 en Ecuador, se identificaron 10 provincias con los índices más significativos de abandono escolar. Morona Santiago encabeza la lista con un 7.3 %, seguida por Esmeraldas con un 7 %, y Pastaza ocupa el tercer lugar con un 6.8 %. Estos datos resaltan la disparidad en los niveles de deserción escolar entre las provincias ecuatorianas con más rezago escolar. Además se evidencia que, en estas provincias, a excepción de Chimborazo, han tenido una reducción en la tasa de deserción entre ambos periodos.

Figura 2.1: Provincias con mayor tasa de abandono escolar periodo 2021 - 2022



Fuente: MINEDUC, 2022
Elaboración propia

Capítulo 3

Metodología

3.1. Datos y población objetivo

Este estudio se basa en datos recabados por la Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU) correspondiente al año 2022, una operación estadística integrada al Sistema Integrado de Encuestas de Hogares (SIEH). El diseño metodológico de la ENEMDU permite analizar la situación del empleo en el país, así como para caracterizar el mercado laboral, la actividad económica, las fuentes de ingresos de la población, así como identificar factores sociodemográficos. Además, la ENEMDU contribuye a nutrir el Sistema de Cuentas Nacionales gestionado por el Banco Central del Ecuador (BCE) (INEC, 2022).

El Instituto Nacional de Empleo (INEM), ente adscrito al Ministerio de Trabajo y Recursos Humanos, ha desempeñado un papel en la gestión de información sobre la dinámica de la fuerza laboral en Ecuador. Iniciando en 1987, el INEM ha llevado a cabo la Encuesta Permanente de Empleo y Desempleo en áreas urbanas de manera anual. En 1993, la responsabilidad de planificación y ejecución de esta encuesta, manteniendo la misma metodología, periodicidad y alcance nacional, urbano y rural, fue transferida al Instituto Nacional de Es-

tadística y Censos (INEC) y a partir de 2003, se adoptó una periodicidad trimestral para la investigación (INEC, 2022).

Por otra parte, el marco teórico subyacente en el cuestionario de las condicones laborales sigue las pautas establecidas por la Conferencia Internacional de Estadísticos del Trabajo (CIET) en 1982, donde la Organización Internacional del Trabajo (OIT) desempeña el papel de secretario técnico. Esta alineación metodológica ha permitido mantener la coherencia y la comparabilidad de los datos recopilados a lo largo del tiempo (INEC, 2022).

La muestra de la Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU) correspondiente al año 2022 contiene información de 336,465 individuos y 27.048 viviendas, lo cual garantiza la representatividad a escala nacional (INEC, 2022b). Para el análisis específico realizado, se ha restringido la muestra a individuos con edades comprendidas entre 5 y 24 años, abarcando aquellos que se encuentran cursando niveles de educación general básica, bachillerato y sistema de educación superior.

En este análisis, la deserción educativa se define como la situación en la que los individuos, que, aunque en el pasado fueron parte del sistema educativo formal, actualmente no asisten a clases, para identificarlos se ha utilizado la variable nivel de instrucción provista por la misma encuesta. La muestra resultante tras aplicar estos criterios se reduce a 122,037 observaciones. En consecuencia, se examinan las circunstancias de las personas en edad de cursar la Educación General Básica (EGB), el Bachillerato General Unificado (BGU) y el nivel superior que están debidamente registradas en el sistema educativo (MINEDUC, 2016). Cabe mencionar, que el análisis de datos faltantes y atípicos se revisará en la siguiente sección.

3.2. Limpieza y preparación de los datos

En esta sección se llevará a cabo el proceso de limpieza y preparación de los datos con el objetivo de garantizar la integridad y calidad de la información utilizada en el análisis (Wooldridge, 2010). El objetivo será identificar posibles problemas, como valores atípicos y datos faltantes, que podrían distorsionar los resultados finales. Para acceder al código fuente de la limpieza y construcción del modelo, visite el repositorio en GitHub: https://github.com/Carolm96/CapstoneUdla.git.

Así, en la Tabla 3.1 se muestra el porcentaje de datos perdidos para cada variable incluida en el modelo. Se observa que la variable correspondiente al ingreso per cápita presenta un 0,38 % de datos faltantes en relación al total de observaciones. Antes de considerar técnicas de imputación, se llevará a cabo un análisis de la estadística descriptiva de las variables y posibles datos atípicos que puedan existir. Este análisis proporcionará una base objetiva para la toma de decisiones respecto a la necesidad y pertinencia de aplicar métodos específicos de imputación de datos (Castro y Ávila, 2006).

Tabla 3.1: Porcentaje de datos perdidos

Variable	Porcentaje del total
Área	0,00 %
Sexo	0,00 %
Edad	0,00 %
RelacionParentesco	0,00 %
EstadoCivil	0,00 %
Etnia	0,00 %
Ingreso per cápita	0,38 %

Fuente: MINEDUC, 2022
Elaboración propia

A continuación, se presenta un resumen de estadística descriptiva de las variables continuas que son la edad y el ingreso per cápita (Ver Tabla 3.2). En cuanto a la edad,

la media de aproximadamente 14.9 años indica que, en promedio, la población tiene una edad cercana a los 15 años. La desviación estándar de alrededor de 5.48 años sugiere una variabilidad moderada en las edades, con un mínimo de 5 años y un máximo de 24 años.

En lo que respecta al Ïngreso Per cápita Familiar", la media es de alrededor de 219.28. Sin embargo, la presencia de un valor mínimo extremadamente bajo (0.29) y un valor máximo muy alto (17.031) indica una gran variabilidad en los ingresos familiares. La desviación estándar de aproximadamente 238.5 refuerza esta observación al indicar una dispersión significativa alrededor de la media. Los percentiles proporcionan información adicional: el 25 % de los ingresos per cápita familiares son inferiores a 88.20, el 50 % son inferiores a 154, y el 75 % son inferiores a 268.

Tabla 3.2: Resumen descriptivo

Estadísticos	Edad	ingr_per
count	122.037	121.573
mean	14,9	219,28
std	5,5	238,5
min	5	0,29
25 %	10	88,20
50 %	15	154
75 %	20	268
max	24	17.031

Fuente: MINEDUC, 2022
Elaboración propia

La desviación estándar de la variable ingreso per cápita sugiere la presencia de valores extremos, lo cual se corrobora en el diagrama de caja y bigotes (Ver Figura A.1 en Anexo A). Para solventar estas desviaciones se recurre a la técnica de rango intercuartílico [3] que permitirá excluir aquellos puntos que son identificados como atípicos (Ver Tabla A.1 en Anexo A). En este caso, el resultado indica que aproximadamente el 6.57 % de los datos

[3]Es una medida estadística de variabilidad que representa la dispersión de la mitad central de un conjunto de datos. Se calcula como la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1) de la siguiente manera [IQR = Q3 - Q1] (Ruiz-Méndez et al., 2020).

en la variable ïngreso per cápita"son considerados extremos (ver Tabla 3.3). Por lo que se procede a excluirlos del conjunto de datos.

Tabla 3.3: Datos atípicos

Métrica	Valor
Total de datos	122.037
Datos atípicos	8.018
Porcentaje atípicos	6,5701 %

Elaboración propia

Una vez removidos los datos atípicos, se obtuvieron nuevamente los estadísticos descriptivos y se procedió a examinar la presencia de valores faltantes en la variable de ingreso per cápita que se obtuvo anteriormente (Ver Tabla A.2 en Anexo A). La muestra resultante, tras la exclusión de los datos extremos, consta de 114,019 observaciones. Este conjunto de datos presenta un porcentaje de datos faltantes del 0.40 % con respecto al total de datos disponibles. Dado que este porcentaje es relativamente bajo, se optará por aplicar la técnica de imputación mediante la sustitución de la media (Romero et al., 2023). En la Tabla 3.4 se muestran las medidas descriptivas finales.

Tabla 3.4: Resumen estadística descriptiva

Estadísticos	Edad	ingr_per
count	114.019	114.019
mean	14,9	173,86
std	5,5	116,8
min	5	0,29
25 %	10	84,75
50 %	15	145
75 %	19	238
max	24	536,25

Fuente: ENEMDU, 2022
Elaboración propia

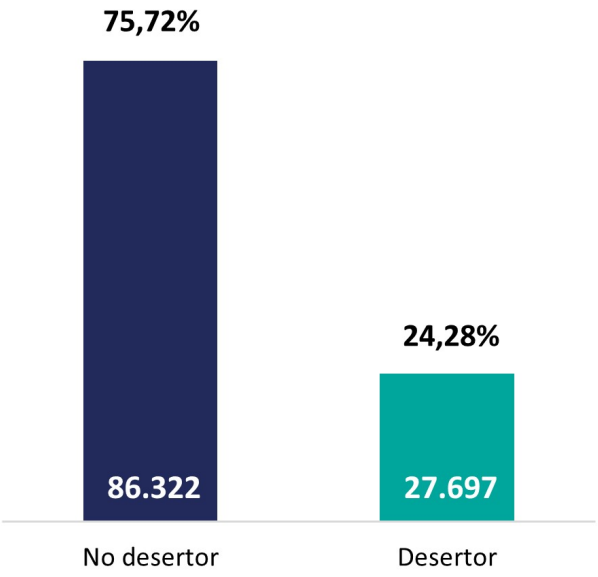
3.3. Descripción de variables

Para la definición de la variable de interés, se recurre a la información provista por la Encuesta Nacional de Empleo, Subempleo y Desempleo (ENEMDU). En este sentido, la variable clave es 'Asiste a clases', que asigna el valor 1 a los individuos que sí asisten y 0 a aquellos ausentes. Sin embargo, existen ciertos niños, niñas y adolescentes que no asisten a clases debido a que nunca ingresaron al sistema educativo, situación que no implica necesariamente una deserción escolar.

Con el propósito de atender esta eventualidad, se incorpora la variable "Nivel de Instrucción"de la encuesta, la cual describe el nivel educativo alcanzado por los individuos en el marco del sistema formal. La primera categoría, denotada como "Ninguno", identifica a aquellos que nunca han accedido al sistema educativo formal, y estas observaciones son excluidas de la investigación. Esto asegura que los individuos que no asisten a clases hayan tenido previamente una vinculación con el sistema educativo formal para que se pueda considerar como deserción. Este enfoque robustece la validez de los resultados al contemplar el historial educativo de los participantes.

En la figura 3.1 se muestra el porcentaje de individuos desertores y no desertores, se puede observar que del total de personas incluidas en el análisis, el 75,72 % corresponde a individuos que permanecen en el sistema educativo, mientras que el 24,28 % son desertores.

Figura 3.1: Frecuencia Deserción Escolar



Fuente: ENEMDU, 2022
Elaboración propia

En esta sección, se presenta la descripción de cada variable de entrada que será empleada en el modelo para predecir la deserción escolar, todas ellas, basadas en la revisión inicial de la literatura. Las características individuales que serán objeto de análisis se detallan en la Tabla 3.1. Estas abarcan tanto aspectos demográficos como socioeconómicos, de naturaleza categóricas y continuas.

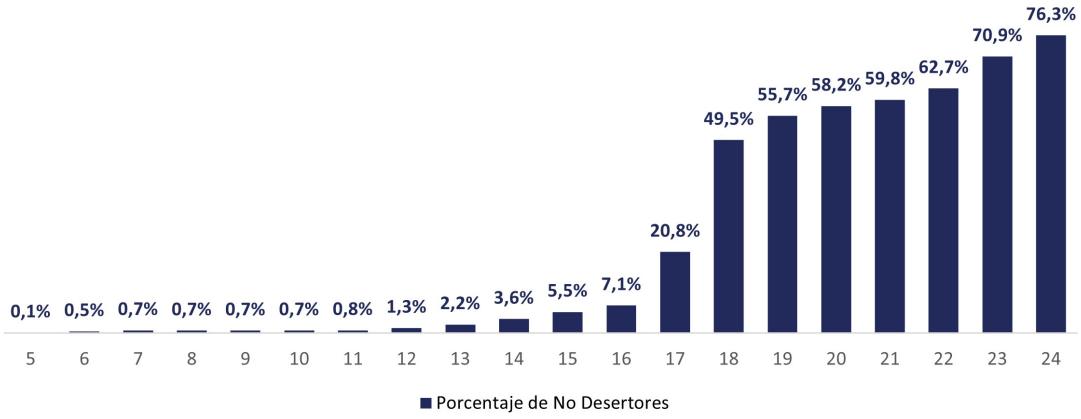
Tabla 3.5: Características individuales

Variable	Descripción	Naturaleza	Autores
Edad	Edad en años del Individuo al momento de realizar la encuesta	Continua	
Área	Variable dicotómica: 1 si el individuo vive en una zona urbana y 0 si reside en una zona rural	Categórica	
Sexo	Variable dicotómica: 1 si la persona es hombre y 0 si es mujer	Categórica	
Estado Civil	1: Casado 2: Separado 3: Divorciado 4: Viudo 5: Unión libre 6: Soltero	Categórica	Román (2013), Valero et al. (2005), Cerón y Verduzco (2005), Cuji et al. (2017), Calero et al. (2023), Ramirez et al. (2022)
Etnia	1: Indígena 2: Afroecuatoriano 3: Negro 4: Mulato 5: Montubio 6: Mestizo 7: Blanco	Categórica	
Relación de parentesco	1: Jefe 2: Cónyuge 3: Hijo/hija 4: Yerno/nuera 5: Nieto/nieta 6: Otros parientes 7: Empleado doméstico 8: Otros no parientes	Categórica	
Ingr_per	Ingreso promedio por miembro del hogar, calculado como la suma total de los ingresos percibidos, dividida entre el número total de integrantes del hogar.	Continua	

Fuente: ENEMDU, 2022
Elaboración propia

En cuanto a la edad, en la Figura 3.2 se observa el porcentaje de estudiantes que desertaron por edad. Las cifras indican que, el mayor porcentaje de desertores se encuentra en la adolescencia y juventud, mientras que, en la niñez el porcentaje es menor.

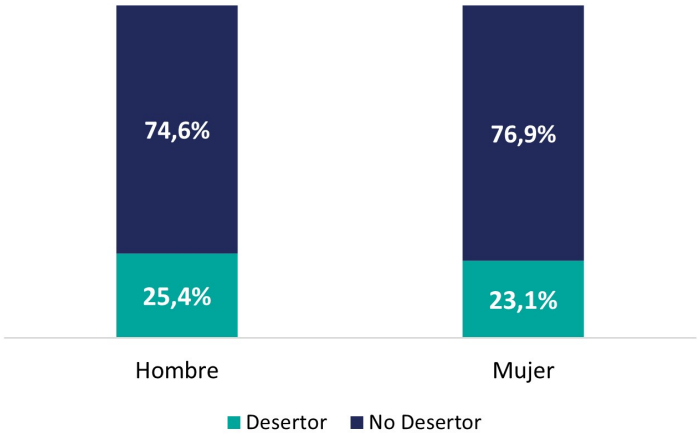
Figura 3.2: Porcentaje de deserción por edad



Fuente: ENEMDU, 2022
Elaboración propia

La Figura 3.3 presenta la distribución porcentual de desertores y no desertores, segmentados por género. Se observa que el 25.4 % de los hombres son desertores, mientras que el 74.6 % no lo son. En comparación, las mujeres muestran un 23.1 % de desertores y un 76.9 % de no desertores. Las mujeres desertan un 2,3 % menos que los hombres.

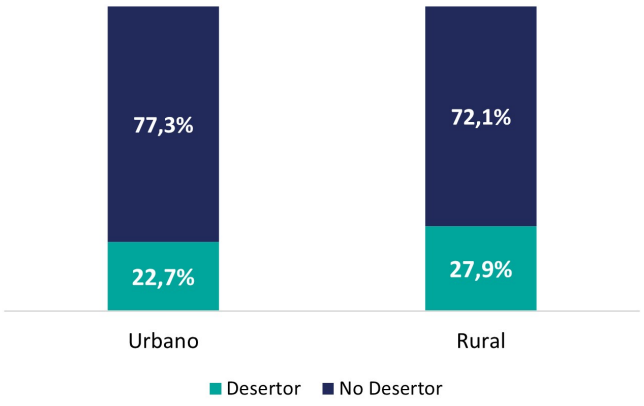
Figura 3.3: Deserción escolar por sexo



Fuente: ENEMDU, 2022
Elaboración propia

En relación a la zona de residencia, el porcentaje de desertores es menor en áreas urbanas (22.7 %) en comparación con áreas rurales (27.9 %). Estos resultados indican que las personas que residen en zonas rurales son más propensas a la deserción escolar que aquellos que residen en zonas urbanas en cerca del 5,2 %.

Figura 3.4: Deserción escolar por zona



Fuente: ENEMDU, 2022
Elaboración propia

En la Tabla 3.6 se observa que aquellas personas que están casadas o en unión libre muestran porcentajes significativamente más altos de deserción, con un promedio del 83,8 % y 84,6 %, respectivamente. Por otro lado, los individuos solteros presentan un porcentaje mucho menor de deserción, con solo un 19,5 %. Estos hallazgos sugieren que el estado civil puede influir en la continuidad educativa, siendo las personas casadas o en unión libre más propensas a interrumpir sus estudios en comparación con los solteros.

En cuanto a la etnia, se destaca que las personas montubias constituyen el porcentaje más elevado de deserción escolar (31,5 %), seguido de las personas indígenas con el 28 %. Por su parte, los mestizos, que son el grupo más numeroso, presentan el porcentaje más bajo de deserción (23,6 %) sugiriendo una mayor estabilidad en la continuidad educativa en comparación con el resto.

Además, la estadística muestra que el cónyuge del jefe de hogar, así como el yerno o la nuera, presentan los índices más altos de deserción del 84,3 % y el 82,1 % respectivamente. Esto sugiere que los miembros de la familia que no son directamente hijos o nietos del jefe de hogar enfrentan mayores desafíos en la continuidad de sus estudios. En comparación, los hijos y nietos del jefe de hogar son menos propensos a abandonar sus estudios, pues presentan cifras más bajas. El resto de categorías de las variables estado civil y etnia que corresponden a la menor proporción de población, así como las personas que no tienen una relación directa con el jefe de hogar (empleado doméstico y otros no parientes) se presentan en la Tabla A.3 de los Anexos.

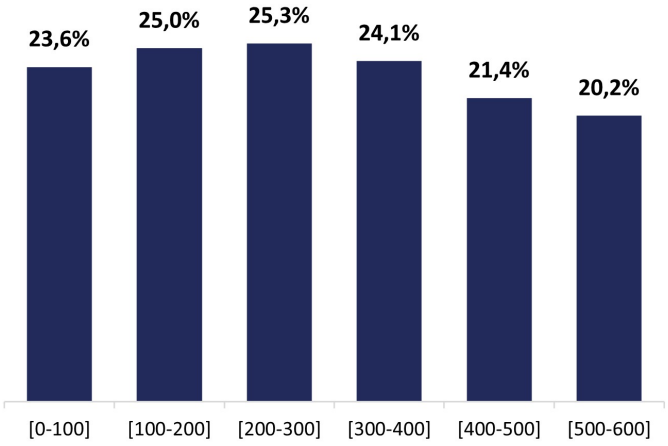
Tabla 3.6: Deserción por estado civil, etnia y relación de parentesco

Variable	Desertor (%)	No Desertor (%)
Estado Civil		
Casado	83,8 %	16,2 %
Separado	85,2 %	14,8 %
Soltero	19,5 %	80,5 %
Unión libre	84,6 %	15,4 %
Etnia		
Indígena	28,0 %	72,0 %
Afroecuatoriano	25,5 %	74,5 %
Negro	27,8 %	72,2 %
Mulato	26,9 %	73,1 %
Montubio	31,5 %	68,5 %
Mestizo	23,6 %	76,4 %
Relación de parentesco con el jefe de hogar		
Jefe	78,7 %	21,3 %
Cónyuge	84,3 %	15,7 %
Hijo/hija	21,9 %	78,1 %
Yerno/nuera	82,1 %	17,9 %
Nieto/nieta	11,9 %	88,1 %
Otros parientes	25,6 %	74,4 %

Elaboración propia

Por último, se observa una tendencia general de disminución en el porcentaje de deserción a medida que aumenta el rango de ingreso per cápita. Los porcentajes más altos de deserción se encuentran en los rangos de ingreso más bajos, con un 23,6 % en el rango [0-100] y un 25,0 % en el rango [100-200]. A medida que aumenta el ingreso per cápita, los porcentajes de deserción tienden a disminuir, alcanzando su punto más bajo en el rango [500-600] con un 20,2 %. Los datos sugieren que el nivel de ingreso per cápita puede tener un impacto positivo en la retención educativa, ya que los porcentajes de deserción disminuyen a medida que aumenta el ingreso.

Figura 3.5: Deserción escolar por rango de ingreso per cápita



Fuente: ENEMDU, 2022
Elaboración propia

3.4. Justificación y aplicación de la Metodología

Varios estudios han abordado la predicción del abandono escolar de los estudiantes mediante el empleo de técnicas de minería de datos como los árboles de decisión, en este contexto, la técnica tiene el propósito de clasificar diversos atributos y anticipar el valor de la clase correspondiente. La estructura de estos árboles comprende nodos y ramas, donde cada nodo representa una condición vinculada a un atributo, y cada rama denota un posible valor

para dicho atributo. Es fundamental destacar que los árboles de decisión tienen la capacidad de manejar tanto valores nominales como numéricos (López et al., 2013; Ramírez et al., 2022 y Márquez et al., 2013).

Entre los algoritmos de árboles de decisión más ampliamente utilizados se encuentran ID3, C4.5 (J48), Naïve Bayes Tree, CART, CHAID. Es importante señalar que el ID3 se limita al procesamiento de datos categóricos y no admite datos numéricos, mientras el algoritmo C4.5 es una extensión que si admite valores numéricos para los atributos (Quinlan, 1996). En este estudio, se utilizará la librería Scikit Learn disponible en Python que se basa en una implementación optimizada del algoritmo CART (Classification and Regression Trees). CART es un algoritmo que puede manejar tanto problemas de clasificación como de regresión (Breiman, 2017).

La presente investigación se centra específicamente en la clasificación de estudiantes con el propósito de determinar la probabilidad de deserción, basándose en atributos y factores diversos. La elección de la técnica de minería de datos recae en el árbol de decisión, motivada por la presencia de variables tanto categóricas como cuantitativas en el conjunto de datos. Camino et al. (2020) respaldan la idoneidad de los árboles de decisión que se fundamenta en su capacidad para manejar distintos tipos de datos y realizar clasificaciones interpretables, lo cual es esencial para comprender y abordar los factores que contribuyen a la deserción estudiantil.

Un árbol de decisión se configura como un conjunto de condiciones organizadas en una estructura jerárquica. Este diagrama incluye un nodo raíz que engloba todas las observaciones, nodos internos que albergan nodos de división, y nodos hoja que contienen las clasificaciones finales para conjuntos específicos de observaciones. Esta técnica representa la segmentación de los datos a través de la aplicación de reglas simples, donde cada regla asigna una observación a un segmento basado en el valor de una entrada. Estas reglas se aplican secuencialmente, creando una jerarquía de segmentos denominada árbol, con cada

segmento llamado nodo. Así, los nodos internos validan atributos, las ramas representan las salidas de las validaciones, y los nodos hoja indican las clases finales asignadas a las observaciones (Márquez et al., 2013).

La técnica de Clasificación Basada en Árboles de Decisión (CBAD), descrita por Song y Ying (2015), se caracteriza por clasificar una población mediante la construcción de un modelo de segmentos dispuestos en forma de árbol invertido. Este modelo se utiliza posteriormente para predecir una variable objetivo y para predecir futuras clasificaciones donde se evalúa cada nuevo caso utilizando este árbol como base de decisiones. Esta estrategia sigue un enfoque «top-down» [3] en una división recursiva (Hofmann y Klinkenberg, 2013). El proceso comienza seleccionando un atributo para el nodo raíz, y se crean ramas para cada posible valor del atributo. Los registros se dividen en subconjuntos según estos valores, repitiendo estos pasos recursivamente para cada rama, utilizando únicamente los registros incluidos en esa rama. El proceso puede detenerse si todos los registros tienen la misma clase.

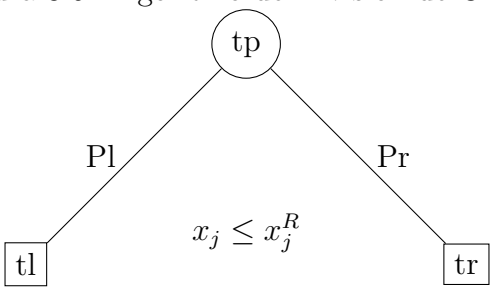
3.4.1. Especificación

Sea tp un nodo padre y tl, tr respectivamente los nodos hijo izquierdo y derecho del nodo padre tp. Se considera la muestra de aprendizaje con la matriz de variables X que tiene M variables x_j y N observaciones. El vector de clases Y consiste en N observaciones con un total de K clases.

La construcción del árbol de clasificación se lleva a cabo siguiendo una regla de división, que es responsable de subdividir la muestra de aprendizaje en partes más pequeñas. Es importante destacar que en cada división, los datos se separan en dos partes con la máxima homogeneidad.

[3] Top-down: Describe la estrategia de construcción del árbol desde la raíz hacia las hojas. Este enfoque implica tomar decisiones sobre qué atributo dividir primero en función de su importancia para la clasificación y, a continuación, realizar divisiones sucesivas en cada nivel del árbol (Hofmann y Klinkenberg, 2013).

Figura 3.6: Algoritmo de División de CART



Fuente: Elaboración propia

Donde tp, tl, tr son los nodos padre, izquierdo y derecho respectivamente; x_j es la variable j; x_j^R es el mejor valor de división de la variable x_j. La homogeneidad máxima de los nodos hijos está definida por la llamada función de impureza $i(t)$. Dado que la impureza del nodo padre tp es constante para cualquiera de las divisiones posibles $x_j \leq x_j^R$, $j = 1, \ldots, M$, la homogeneidad máxima de los nodos hijos izquierdo y derecho será equivalente a la maximización del cambio de la función de impureza $\Delta i(t)$:

$$\Delta i(t) = i(tp) - E[i(tc)]$$

donde tc son los nodos hijos izquierdo y derecho del nodo padre tp. Suponiendo que P_{li} y P_{ri} son las probabilidades de los nodos izquierdo y derecho, se obtiene:

$$\Delta i(t) = i(tp) - P_{li} i(tl) - P_{ri} i(tr)$$

Por lo tanto, en cada nodo, CART resuelve el siguiente problema de maximización:

$$\arg \max_{x_j \leq x_j^R, j=1,\ldots,M} [i(tp) - P_{li} i(tl) - P_{ri} i(tr)] \tag{3.1}$$

La Ecuación 3.1 implica que CART buscará a través de todos los posibles valores de todas las variables en la matriz X para la mejor pregunta de división $x_j < x_j^R$ que maximizará el cambio en la medida de impureza $\Delta i(t)$ (Timofeev, 2004).

Una interrogante crucial consiste en la definición de la función de impureza $i(t)$. Una de las más ampliamente empleadas en la práctica es la entropía [4]. Esta regla desempeña un papel fundamental en la evaluación de la calidad de las divisiones en los nodos del árbol de decisión.

Con el propósito de realizar un análisis predictivo, se procedió a realizar una división aleatoria del conjunto de datos en dos submuestras. La primera submuestra, que comprende el 75 % de los registros, fue asignada como el conjunto de entrenamiento. La segunda submuestra, constituida por el 25 % restante de los registros, se reservó para evaluar la capacidad predictiva del modelo como el conjunto de prueba. Además, Se utilizará el criterio de entropía para medir la impureza de los nodos durante la construcción del árbol.

Para el entrenamiento del modelo, se empleó la clase DecisionTreeClassifier de la biblioteca scikit-learn, se configuraron los parámetros del modelo, como el criterio de división y la profundidad máxima del árbol (5 niveles). El modelo fue ajustado utilizando el conjunto de entrenamiento X_train y y_train mediante el método fit. Durante este proceso, el Árbol de Decisión aprendió las relaciones entre las características de entrada X_train y las etiquetas de salida y_train, utilizando el criterio de entropía para medir la impureza de los nodos. En el capítulo siguiente se discutirá el rendimiento del modelo.

[4]Es un criterio para medir la impureza de un conjunto de datos en un nodo del árbol. La entropía evalúa la homogeneidad de una colección de datos en un nodo. Cuanto menor sea la entropía, más homogéneo será el conjunto de datos en términos de clases.

Capítulo 4

Discusión de Resultados

En este capítulo se presentan los resultados obtenidos de la aplicación del modelo. En la Tabla 4.1 se muestra la matriz de confusión que refleja la evaluación del modelo en un conjunto de datos de prueba utilizado para clasificar la deserción escolar. En esta, se observa que hay 18.859 verdaderos negativos, 2.773 falsos positivos, 1.588 falsos negativos y 5.285 verdaderos positivos. Este análisis revela que el modelo tiende a clasificar correctamente a los estudiantes no desertores, pero muestra cierta dificultad en identificar de manera precisa a aquellos que realmente abandonan la escuela.

Tabla 4.1: Matriz de Confusión

		Predicción	
Valor Real	Clase	**No desertaron**	**Desertaron**
	No desertaron	18.859	2.773
	Desertaron	1.588	5.285

Elaboración propia

En la Tabla 4.2 se presentan métricas cuantitativas que brindan una perspectiva del desempeño del modelo en la tarea de clasificación de la deserción escolar. Estas métricas

incluyen precision, recall, f1-score y support, cada una aportando información específica sobre diferentes aspectos del rendimiento del modelo.

Los resultados de desempeño del modelo revelan una capacidad destacada para clasificar correctamente a los estudiantes no desertores, con una precisión del 92 %, indicando que el 92 % de las predicciones positivas para esta clase son correctas. Asimismo, se observa un recall del 87 % para esta misma clase, lo que indica que el modelo identifica eficazmente al 87 % de los estudiantes no desertores en el conjunto de datos. Sin embargo, el rendimiento es menor en la clasificación de desertores, con una precision del 66 % y un recall del 77 %, sugiriendo cierta dificultad para identificar con precisión a los estudiantes que realmente abandonan la escuela.

Para mejorar la precisión del 66 % en la clasificación de desertores, es crucial abordar las posibles causas subyacentes de esta limitación. Una solución sería la recopilación de nuevas variables relevantes como el rendimiento académico. Esto podría permitir al modelo capturar mejor la complejidad de los factores que influyen en la deserción escolar. Además, la aplicación de técnicas, como el ajuste de hiperparámetros y la exploración de algoritmos más complejos, podría ayudar a mejorar la capacidad del modelo para discernir patrones y mejorar su rendimiento predictivo en la identificación de desertores.

La exactitud general del modelo es del 85 %, reflejando su capacidad para clasificar correctamente la mayoría de los casos en el conjunto de datos de prueba, mientras que las métricas macro y ponderada proporcionan una visión equilibrada del rendimiento general del modelo en ambas clases. Estos hallazgos son respaldados por la evaluación del modelo en un conjunto de datos de 28,505 instancias, siendo 21,632 no desertores y 6,873 desertores.

Tabla 4.2: Resultados de Desempeño del Modelo

	Precision	Recall	F1-Score	Support
Clase 0	0.92	0.87	0.90	21.632
Clase 1	0.66	0.77	0.71	6.873
Accuracy	0.85			
Macro Avg	0.79	0.82	0.80	28.505
Weighted Avg	0.86	0.85	0.85	28.505

Elaboración propia

A pesar de la capacidad del modelo para clasificar correctamente a los estudiantes no desertores, es importante evaluar la presencia de sobreajuste que ocurre cuando el modelo se ajusta demasiado a los detalles del conjunto de entrenamiento y, en consecuencia, no generaliza bien a nuevos datos.

Para abordar esta inquietud, se llevaron a cabo análisis durante la fase de construcción del modelo. Se aplicaron técnicas destinadas a evaluar el sobreajuste, como el análisis de validación cruzada y la examinación de curvas AUC-ROC. Los resultados obtenidos revelan que no hay evidencia de sobreajuste en el modelo (consultar Tabla B.1 y Figura B.1 en el Anexo B).

El árbol de decisión construido para predecir la deserción escolar presenta múltiples niveles de división basados en variables como la edad, estado civil y sexo, etc. En el nodo raíz, se establece la primera división en función de la edad, donde aquellos menores o iguales a 17 años son mayormente clasificados como no desertores. La rama derecha del nodo de estado civil revela que, entre aquellos que no cumplen con la condición inicial de edad, aquellos que están casados, separados, divorciados son predominantemente clasificados como desertores. En este mismo nodo se analiza nuevamente la edad, si la edad es menor o igual a 19.5 años, se observa una mayor probabilidad de deserción, especialmente para aquellos menores de 18.5 años.

Dentro del nodo de edad menor o igual a 18.5 años, la variable de ingreso per cápita se utiliza para continuar con la clasificación. Se destaca que aquellos con ingresos per cápita menores o iguales a $352.6 muestran una mayor propensión a la deserción (26,6 %) en comparación con aquellos cuyos ingresos per cápita son menores o iguales a $136 (18,5 %). Este hallazgo sugiere que el nivel socioeconómico, representado por el ingreso per cápita, influye en la decisión de deserción escolar en este grupo demográfico.

Por otro lado, en el nodo donde se evalúa si el sexo es menor o igual a 0.5 (mujer). Aquí, se observa que la rama derecha de esta condición está asociada con una clasificación positiva, indicando una mayor propensión a la deserción para las mujeres. A medida que se profundiza en este nodo, se introduce la variable de ingreso per cápita, destacando que ingresos más bajos, específicamente menores o iguales a $154.5, se asocian con una mayor probabilidad de deserción.

Cuando una persona es soltera, reside en una zona rural, es de sexo femenino y tiene una edad menor a 22 años implicaría una mayor probabilidad de ser desertores. Por el contrario, cuando una mujer reside en zonas urbanas, es soltera y tiene menos de 22 años se clasificaría como no desertora. Asimismo, cuando una persona reside en zonas urbanas, tiene más de 22 años pero tiene un ingreso per cápita menor a $212 se clasificaría como desertora, denotando una vez más que los bajos ingresos implicarían en las decisiones de abandonar la escuela.

En la rama de individuos menores de 17 años, la segmentación por edad resalta la influencia diferencial de ingresos per cápita. Se observa que ingresos per cápita inferiores a $91.05 están asociados con tasas bajas de deserción en este grupo específico. Esta relación inversa entre ingresos y deserción indica que, en edades tempranas, los bajos ingresos no limitan por sí solos la deserción, sino que puede deberse a otros factores también.

En el nodo donde la relación de parentesco es menor o igual a 6, englobando roles como jefe de hogar, conyuge, hijos, se observa una menor entropía, indicando que la relación entre estos roles familiares y la deserción escolar es más predecible. Este hallazgo sugiere que, los individuos con roles familiares más cercanos y directos tienen patrones más consistentes en cuanto a la decisión de permanecer en la educación.

En cuanto a la etnia, se observa que la mayoría de las muestras (2.775) pertenecen a la clase (no desertores), mientras que una minoría (160) pertenece a la clase (desertores). Esto indica que, en este subconjunto específico caracterizado por ingreso per cápita y etnia, la variable etnia está siendo utilizada para clasificar predominantemente a los individuos como no desertores.

En resumen, cuando la edad supera los 17 años, el modelo realiza divisiones basadas en el estado civil, la zona de residencia, el sexo, y el ingreso per cápita. Destaca la influencia del ingreso per cápita, dividiendo a los individuos en subgrupos según este criterio. Específicamente, se observa una tendencia a clasificar a aquellos con ingresos más bajos como desertores. Por otro lado, cuando los individuos son menore a 17 años, la rama se divide en subconjuntos según la edad y el ingreso per cápita. Se destaca un patrón donde ingresos per cápita inferiores a $91.05 están asociados con tasas bajas de deserción, sugiriendo que, en edades tempranas, ingresos más bajos no actúan como un único factor que influye en la deserción. Además, se observa cómo la relación de parentesco y la etnia influyen en la clasificación. Roles familiares más cercanos y directos, como jefe de hogar, conyuge, hijos, tienden a presentar patrones más consistentes en la decisión de permanecer en el sistema educativo.

4.1. Diseño de Estrategias

Considerando los resultados del modelo de árbol de decisión, donde se evidencia la relevancia de variables como la edad y el ingreso per cápita en la predicción de la deserción escolar, se podría sugerir una estrategia continua de monitoreo y evaluación del riesgo de deserción para cada estudiante. Es decir, establecer un sistema automatizado que, utilizando los datos demográficos y socioeconómicos de cada estudiante, ejecute el modelo de árbol de decisión periódicamente. Esto permitirá obtener una evaluación actualizada del riesgo de deserción para cada individuo.

Dentro de este sistema, se establecerían alertas para identificar a los estudiantes que muestran un aumento en su riesgo de deserción. Por ejemplo, si un estudiante muestra cambios significativos en variables como el ingreso per cápita que lo colocan en una categoría de mayor riesgo, se activaría una alerta para que el personal escolar pueda intervenir de manera oportuna. Al analizar los resultados del modelo, se observa que aproximadamente el 24 % de los estudiantes son clasificados como desertores. Esta cifra representa una parte significativa de la población estudiantil y destaca la importancia de implementar medidas preventivas y de intervención dirigidas específicamente a este grupo.

Además de la implementación del sistema automatizado, es importante evaluar su efectividad en la reducción de la deserción escolar. Para ello, se podrían establecer métricas de éxito, como la tasa de deserción escolar antes y después de la implementación del sistema. También, se pueden considerar indicadores adicionales, como la mejora en el rendimiento académico de los estudiantes identificados como en riesgo. Estos criterios permitirán una evaluación exhaustiva del impacto del sistema y proporcionarán información valiosa para ajustes y mejoras continuas.

Por otra parte, dada la significativa proporción de estudiantes que residen en áreas rurales y que el modelo los clasifica como desertores, es relevante implementar estrategias efectivas para abordar este desafío específico. Por lo tanto, se sugiere una estrategia de implementación a nivel provincial haciendo una distinción entre zonas urbanas y rurales. Según los datos obtenidos, el 55 % de los estudiantes que residen en áreas rurales son clasificados como desertores por el modelo, lo que subraya la importancia de dirigir esfuerzos hacia esta población en particular.

Se puede proponer estrategias junto con las autoridades educativas para adaptar y replicar el modelo en diferentes regiones, provincias y cantones del país. Cada provincia podría ajustar el modelo a sus necesidades y contextos específicos, incorporando variables adicionales relevantes para la región. Factores como el acceso a servicios educativos, la disponibilidad de transporte escolar o las condiciones económicas locales podrían ser considerados en el modelo para mejorar su precisión en cada área.

El establecimiento de este modelo a nivel provincial permitiría una identificación más precisa de los estudiantes en riesgo de deserción en cada región. Así, se podrían diseñar intervenciones más efectivas y dirigidas, contribuyendo a reducir la deserción escolar y promover un ambiente educativo más inclusivo y equitativo.

4.2. Implicaciones sobre Innovación

El uso de modelos predictivos como el árbol de decisión en el ámbito educativo presenta una oportunidad para innovar en la manera en que se abordan los desafíos relacionados con la deserción escolar. Al permitir una identificación temprana de los estudiantes en riesgo de abandonar la escuela, estos modelos brindan una herramienta eficaz para intervenir de

manera proactiva y personalizada. Esta capacidad de anticiparse a los problemas y tomar medidas preventivas representa un enfoque innovador en la gestión educativa.

La implementación de intervenciones personalizadas y programas de apoyo basados en los resultados de estos modelos también constituye una innovación en la forma en que se aborda la retención estudiantil. En lugar de enfoques estandarizados, las acciones se adaptan específicamente a las necesidades y circunstancias de cada estudiante, lo que aumenta la probabilidad de éxito y el impacto positivo en su trayectoria académica.

Sin embargo, es importante reconocer la posibilidad de sesgos en el modelo para identificar estudiantes en riesgo de deserción escolar. Estos sesgos pueden manifestarse en los datos utilizados para entrenar los algoritmos o en el propio diseño del modelo, lo que podría resultar en decisiones injustas o inequitativas.

Esta innovación en la gestión educativa no solo beneficia a los estudiantes y al sistema escolar, sino que también puede tener implicaciones positivas para el sector empresarial y la sociedad en general. Al reducir la deserción escolar y mejorar las tasas de retención estudiantil, se crea un conjunto de talento más sólido y calificado, lo que a su vez puede beneficiar a las empresas al proporcionarles acceso a una fuerza laboral más educada y capacitada.

Por otra parte, al priorizar los recursos y esfuerzos en los estudiantes identificados como en riesgo de deserción escolar, el modelo ayudaría a las instituciones educativas a utilizar sus recursos de manera más eficiente y efectiva. Esto puede conducir a una mejor asignación de presupuestos y personal, maximizando el impacto de las iniciativas de retención estudiantil. Además, se puede contribuir a mejorar la equidad educativa, al proporcionar apoyo adicional a los estudiantes en situaciones de riesgo, se reduce la disparidad en el acceso a la educación y se promueven la igualdad de oportunidades para todos.

Por último, el diseño de políticas públicas podría estar dirigido a mejorar la situación económica de las familias. Al identificar cómo el ingreso per cápita de las familias afecta significativamente la probabilidad de deserción escolar, el modelo proporciona información para la formulación de políticas económicas y sociales orientadas a reducir esta brecha. Por ejemplo, los gobiernos podrían desarrollar programas de asistencia financiera dirigidos específicamente a familias con ingresos bajos o inestables, promoviendo así la igualdad de oportunidades económicas.

Capítulo 5

Conclusiones y Recomendaciones

Este estudio analizó e identificó los factores demográficos y socioeconómicos que clasifican a un estudiante en riesgo de desertar o no del sistema educativo. La deserción escolar fue seleccionada para el análisis debido a que, según cifras estadísticas del INEC, en el periodo escolar 2021-2022, la tasa de abandono escolar se ubicaba en 2,11 %, cifra que ha aumentado en comparación con los últimos tres periodos escolares anteriores. Esto denota que niños, jóvenes y adolescentes han decidido interrumpir sus estudios por diversas razones, lo que representa un problema tanto a nivel individual como social.

Entre los resultados obtenidos se observa una influencia de variables demográficas como la edad, estado civil, sexo y zona de residencia en las predicciones de deserción escolar. Estas variables resultan importantes para comprender los patrones de deserción en la población estudiantil. Los hallazgos sugieren que el ingreso per cápita juega un papel significativo en las decisiones de deserción escolar, especialmente entre los grupos demográficos más jóvenes. Se encontró una asociación directa entre ingresos más bajos y una mayor probabilidad de deserción, resaltando la importancia de considerar el contexto socioeconómico al abordar la deserción escolar.

Se identificó una disparidad de género en las tasas de deserción, con una mayor propensión a la deserción entre las mujeres en ciertos grupos demográficos. Por otra parte, la relación de parentesco también emergió como un factor relevante, mostrando que los roles familiares más cercanos y directos están asociados con patrones más consistentes en la decisión de permanecer en la educación, subrayando la influencia del apoyo familiar en la retención escolar de los estudiantes.

En este sentido, la aplicación de técnicas de minería de datos supervisadas es fundamental para identificar patrones y factores que influyen en la deserción escolar, permitiendo clasificar a los individuos según su propensión a abandonar el sistema educativo. Una vez que se construye el árbol de decisión, se puede utilizar para predecir la probabilidad de deserción para nuevos casos, asignando a cada caso una clasificación basada en los caminos que siguen a través del árbol. Esto permite a las instituciones educativas y a los responsables de políticas identificar de manera precisa y temprana a los estudiantes en riesgo de deserción, facilitando la implementación de intervenciones específicas y dirigidas.

Se recomienda establecer un proceso de evaluación continua del modelo utilizado para predecir la deserción escolar. Esta evaluación permitirá identificar posibles áreas de mejora y ajustes, además, se debe realizar una evaluación constante de las variables utilizadas en el modelo predictivo para asegurar su relevancia y validez en el contexto específico de la deserción escolar en el Ecuador. Es importante estar atento a posibles cambios en el entorno socioeconómico y demográfico que puedan afectar la capacidad predictiva del modelo.

Es importante la actualización y mejora continua del modelo, esto implicaría la inclusión de nuevas variables basadas en la revisión de literatura que no se han abordado en esta investigación por falta de disponibilidad de la información, ajustes en el algoritmo o la optimización de los hiperparámetros y la división aleatoria de las submuestras. El monitoreo de los indicadores de desempeño como especificidad, sensibilidad, etc. También permitirán evaluar la efectividad del modelo y mejorar la precisión.

Referencias

Becker, G. S. (1962). Investment in human capital: A theoretical analysis. *Journal of political economy*, 70:9–49.

Beyer, H. (1998). ¿ desempleo juvenil o un problema de deserción escolar? *Estudios públicos*, (71):89–119.

Breiman, L. (2017). *Classification and regression trees*. Routledge.

Calero, R., Sosa, M., Lino, L., and Ponciano, J. (2023). Factores determinantes de la deserción escolar en la región huánuco, perú. *Desafíos*, 14(2):118–123.

Camino, J., Urbina, A., and Barbosa, R. (2020). Deserción escolar universitaria: Patrones para prevenirla aplicando minería de datos educativa. *RELIEVE. Revista Electrónica de Investigación y Evaluación Educativa*, 26(1).

Castillo, D., Óscar Espinoza, González, L., and Loyola, J. (2014). Factores familiares asociados a la deserción escolar en los niños y niñas mapuche: un estudio de caso. *Estudios Pedagógicos*, 40:97–112.

Castro, L. M. U. and Ávila, D. M. M. (2006). Una introducción a la imputación de valores perdidos. *Terra. Nueva Etapa*, 22(31):127–151.

Cerón, H. and Verduzco, M. (2017). Aplicación de las técnicas árboles de decisión y redes neuronales para la generación del modelo para la proyección de la deserción escolar. *TecnoCultura*, pages 8–8.

Constitución del Ecuador (2008). Sección quinta: Educación. Technical report, Asamblea Constituyente, Montecristi - Ecuador.

Coronel, A. (2010). Capacitación del capital humano como una inversión para desarrollo. *EUREKA*, 7:71–76.

Cuji, B., Gavilanes, W., and Sanchez, R. (2017). Modelo predictivo de deserción estudiantil basado en arboles de decisión. *Espacios*, 38(55):17.

Eckert, K. B. and Suénaga, R. (2015). Análisis de deserción-permanencia de estudiantes universitarios utilizando técnica de clasificación en minería de datos. *Formación universitaria*, 8(5):03–12.

Espinoza, Ó., Castillo, D., González, L. E., Loyola, J., Cruz, S., et al. (2014). Deserción escolar en chile: un estudio de caso en relación con factores intraescolares. *Educación y Educadores*, 17(1):32–50.

Espínola, V. and Claro, J. (2010). Estrategias de prevención de la deserción en la educación secundaria: perspectiva latinoamericana. *Revista de Educación*, pages 257–280.

García, J., Oropeza, J., and Ordoñez, M. (2012). El teorema de bayes como una herramienta en la mejora de calidad del servicio educativo. *Ava Cient*, 8(2):123–137.

Gómez, C., Padilla, A., and Rincón, C. (2016). Deserción escolar de adolescentes a partir de un estudio de corte transversal: Encuesta nacional de salud mental colombia 2015. *Revista Colombiana de Psiquiatría*, 45:105–112.

Hofmann, M. and Klinkenberg, R. (2016). *RapidMiner: Data mining use cases and business analytics applications*. CRC Press.

Instituto Nacional de Estadística y Censos (INEC) (2021). Condiciones de Vida según nivel de Preparación Académica. Boletín técnico, Dirección de Estadísticas Económicas, Quito - Ecuador.

Instituto Nacional de Estadística y Censos (INEC) (2022a). Encuesta Nacional de Empleo, Desempleo y Subempleo ENEMDU Anual. Diseño muestral, Dirección de Infraestruc-

tura Estadística y Muestreo (DINEM), Quito - Ecuador.

Instituto Nacional de Estadística y Censos (INEC) (2022b). Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU), anual 2022. Technical report, Dirección de Estadísticas Sociodemográficas, Quito - Ecuador.

Levin, B. (2003). Approaches to equity in policy for lifelong learning. Technical report, Organización para la Cooperación y Desarrollo Económico (OCDE).

López, M., Pérez, J., Escobar, Saturnino, M., and Víctor, L. (2013). Análisis comparativo de algoritmos de minería de datos para predecir la deserción escolar. *Advances in Computing Science*, 67:13–23.

Lugo, B. (2013). La deserción estudiantil: ¿realmente es un problema social? *Revista de Postgrado FACE-UC*, 7:289–309.

Márquez, C., Romero, C., and Ventura, S. (2013). Predicting school failure and dropout by using data mining techniques. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, 8(1):7–14.

Melis, F., Díaz, R., and Palma, A. (2005). Adolescentes y jóvenes que abandonan sus estudios antes de finalizar la enseñanza media: Principales tendencias. *División Social MIDEPLAN, Santiago, Chile*.

Ministerio de Educación (MINEDUC) (2016). Acuerdo-ministerial-nro.-mineduc-me-2016-00020-a. Technical report, Quito - Ecuador.

Ministerio de Educación (MINEDUC) (2022). Reporte anual de información educativa. Technical report, Ministerio de Educación, Quito - Ecuador.

Moreno, D. and Moreno, A. (2013). La deserción escolar: Un problema de carácter social. *Revista Internacional de Psicología*, 6:115–124.

Nilsson, N. J. (2005). Introduction to machine learning. an early draft of a proposed textbook.

Ocampo, J. A. and CEPAL, N. (2000). *Equidad, desarrollo y ciudadanía*. Alfaomega.

Pérez, B., Castellanos, C., and Correal, D. (2018). Predicting student drop-out rates using data mining techniques: A case study. In *IEEE Colombian Conference on Applications in Computational Intelligence*, pages 111–125. Springer.

Quinlan, J. R. (1996). Improved use of continuous attributes in c4. 5. *Journal of artificial intelligence research*, 4:77–90.

Ramírez, A., Jiménez, S., Marquez, B., and Martínez-Ramírez, Y. (2022). Un modelo de minería de datos para predecir la deserción escolar en la carrera de ingeniería de software a data mining model to predict school dropout in the software engineering career. *Abstraction Application*, 36:46–62.

Rochin, F. (2021). Deserción escolar en la educación superior en méxico: revisión de literatura. *Revista Iberoamericana para la Investigación y el Desarrollo Educativo RIDE*, 12.

Romero, G., González, C., Díaz, M., and Rueda, N. (2023). Revisión y perspectivas para la construcción de bases de datos robustas con datos faltantes: caso aplicado a información financiera. *Tecnura*, 27(75):14–37.

Román, M. (2013). Factores asociados al abandono y la desercIÓn escolar en amÉrica latina: Una mirada en conjunto. *Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación REICE*, 11:33–59.

Rossel, C., Manzi, P., Antía, F., and Atuesta, B. (2022). Transferencias monetarias no contributivas y educación: Impacto y aprendizajes. Technical report, Comisión Económica para América Latina y el Caribe (CEPAL).

Ruiz-Méndez, D., Quezada, M. E., and Valero, C. Z. V. (2020). Estadística robusta aplicada a las medidas de localización y escala: Nota técnica. *Revista Digital Internacional de Psicología y Ciencia Social*, 6(2):499–517.

Salazar, C. (2017). Deserción escolar. *Con-Ciencia Boletín Científico de la Escuela Preparatoria*, 4.

Sapelli, C. and Torche, A. (2004). Deserción escolar y trabajo juvenil: ¿dos caras de una misma decisión? *CUADERNOS DE ECONOMIA*, 41:173–198.

Silvera, L. (2016). La evaluación y su incidencia en la deserción escolar: ¿falla de un sistema, de las instituciones educativas, del docente o del estudiante? *Educación y Humanismo*, 18:313–325.

Solís, J., Quiroz, S., and Fosado, O. (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*, 7:1–14.

Song, Y.-Y. and Ying, L. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2):130.

Tarupi, E. (2015). El capital humano y los retornos a la educación en ecuador. *Estudios de la Gestión: Revista Internacional de Administración*, pages 81–94.

Timofeev, R. (2004). Classification and regression trees (cart) theory and applications. *Humboldt University, Berlin*, 54.

UNESCO, Declaración de Incheon (2011). La unesco y la educación:"toda persona tiene derecho a la educación". *Francia*.

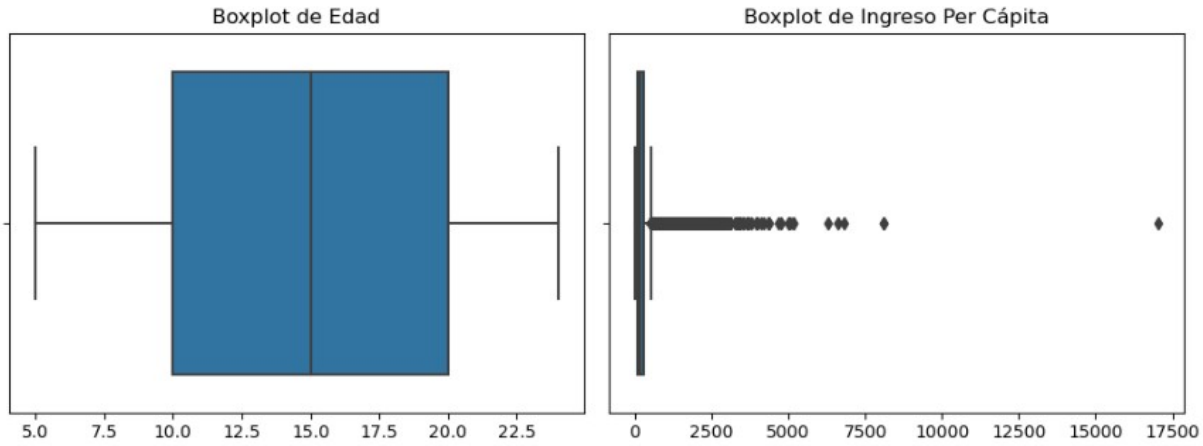
Valero, S., Salvador, A., and García, M. (2005). Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. *Ene*, 779(73):33.

Wooldridge, J. M. (2010). *Introducción a la econometría. Un enfoque moderno*. Cengage Learning, México, 4ª edition.

Anexo A

Limpieza de datos

Figura A.1: Diagrama de caja y bigotes



Fuente: ENEMDU, 2022

Elaboración propia

El rango intercuartílico proporciona una medida de dispersión que se utiliza para identificar y remover valores atípicos. En este caso, los datos atípicos son aquellos que caen fuera de los límites establecidos por 1.5 veces el IQR por debajo del primer cuartil y por encima del tercer cuartil (Ver Tabla A.1).

Tabla A.1: Rango Intercuartílico

Parámetros	Definición	Valor
Primer cuartil Q1:	Represente el 25 % de los datos	88,33
Tercer cuartil Q3:	Corresponde al 75 % de los datos	267,50
Rango intercuartílico=	Q3 - Q1	179,17
Limite inferior =	Q1 - 1.5 * IQR	-180,42
Límite superior =	Q3 + 1.5 * IQR	536,25

Fuente: ENEMDU, 2022
Elaboración propia

Al eliminar los valores extremos en la sección 3.2 del Capítulo 3, la muestra resultante corresponde a 114.019 obervaciones, sin embargo, la variable ingreso per cápita contenía valores nulos, por lo que se procedió a imputar estos datos mediante la sustitución de la media (174,04) (Ver Tabla A.2).

Tabla A.2: Resumen descriptivo

Estadísticos	Edad	ingr_per
count	114.019	114.019
mean	14,9	174,04
std	5,5	116,8
min	5	0,29
25 %	10	84,75
50 %	15	145
75 %	19	238
max	24	536,25

Nota: Resumen descriptivo, una vez que se removieron los datos extremos.
Fuente: ENEMDU, 2022
Elaboración propia

Ahora, se analiza la matriz de correlación (Tabla A.3), la cual muestra que hay una correlación débil y positiva entre la variable edad e ingreso per cápita. Este valor cercano a cero (0,0965) indica que el cambio en una variable no se asocia fuertemente con el cambio en la otra variable

Tabla A.3: Matriz de Correlación

Variable	Edad	ingr_per
Edad	1,000	0,097
ingr_per	0,097	1,000

Fuente: ENEMDU, 2022
Elaboración propia

Tabla A.4: Deserción escolar por estado civil, relación de parentesco y etnia

Variable	Desertor (%)	No Desertor (%)
EstadoCivil		
Viudo	64,0 %	36,0 %
Divorciado	68,8 %	31,3 %
Etnia		
Blanco	31,8 %	68,2 %
Relación de parentesco con el jefe de hogar		
Empleado doméstico	7,2 %	92,8 %
Otros no parientes	67,4 %	32,6 %

Fuente: ENEMDU, 2022
Elaboración propia

Anexo B

Evaluación de Sobreajuste

En el análisis de validación cruzada, se observa que la precisión promedio en el conjunto de entrenamiento es de 0.851, mientras que en el conjunto de prueba es de 0.849. La consistencia entre estas precisiones y la estabilidad a lo largo de las iteraciones sugieren que el modelo no exhibe un sobreajuste significativo. La diferencia marginal entre las precisiones en ambos conjuntos, así como la variabilidad limitada, indican que el modelo generaliza de manera efectiva a datos no vistos (Tabla B.1).

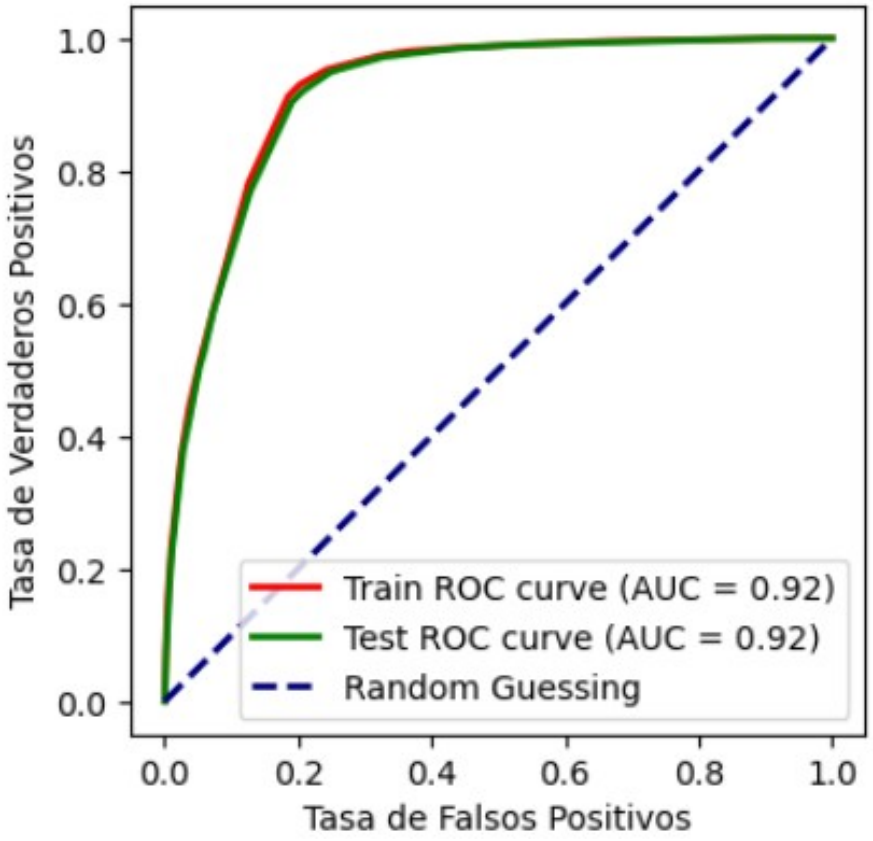
Tabla B.1: Validación Cruzada

	Precisión	
Iteración	**Conjunto de Entrenamiento**	**Conjunto de Prueba**
1	0,854	0,845
2	0,850	0,850
3	0,849	0,849
4	0,849	0,849
5	0,853	0,853
Promedio	**0,851**	**0,849**

Elaboración propia

Complementando lo anterior, se procedió a evaluar el rendimiento del modelo mediante la construcción de la curva ROC en los conjuntos de entrenamiento y prueba, revelando un AUC de 0.92 en ambos casos. La consistencia en los valores de AUC sugiere que el modelo mantiene su capacidad de discriminación entre clases en ambos conjuntos. La coherencia entre la precisión promedio y los resultados de la curva ROC respalda la conclusión anterior de que no se observan signos significativos de sobreajuste (Figura B.1).

Figura B.1: Curva ROC



Elaboración propia

Figura B.2: Árbol de Decisión