UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

IDENTIFICACIÓN DE FACTORES DE DESERCIÓN UNIVERSITARIA MEDIANTE SELECCIÓN DE CARACTERÍSTICAS EN MODELOS DE APRENDIZAJE AUTOMÁTICO

TESIS PARA OBTENER EL TÍTULO DE:

MAESTRO EN TECNOLOGÍAS PARA EL APRENDIZAJE Y EL CONOCIMIENTO

PRESENTA:

L.C.C. DANIEL DOMÍNGUEZ GÓMEZ

BAJO LA DIRECCIÓN DE:

DR. JUAN DE DIOS GONZÁLEZ TORRES

BAJO LA CODIRECCIÓN DE

DR. ARTURO CORONA FERREIRA

Cunduacán, Tabasco. **Agosto, 2025.**

Declaración de Autoría y Originalidad

En la ciudad de Cunduacán, el día 07 del mes Agosto del 2025, el que suscribe Daniel Domínguez Gómez alumna(o) del Programa de Maestría, con número de matrícula 231H20007 adscrito a la Maestría en Tecnologías para el Aprendizaje y el Conocimiento (MTAC), de la Universidad Juárez Autónoma de Tabasco (UJAT), como autor de la Tesis presentada para la obtención del titulo y titulada IDENTIFICACIÓN DE FACTORES DE DESERCIÓN UNIVERSITARIA MEDIANTE SELECCIÓN DE CARACTERÍSTICAS EN MODELOS DE APRENDIZAJE AUTOMÁTICO dirigida por el Dr. Juan de Dios González Torres en codirección del Dr. Arturo Corona Ferreira

DECLARO QUE:

La Tesis es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, de acuerdo con el ordenamiento jurídico vigente, en particular, la LEY FEDERAL DEL DERECHO DE AUTOR (Decreto por el que se reforman y adicionan diversas disposiciones de la Ley Federal del Derecho de Autor del 01 de Julio de 2020 regularizando y aclarando y armonizando las disposiciones legales vigentes sobre la materia), en particular, las disposiciones referidas al derecho de cita.

Del mismo modo, asumo frente a la Universidad cualquier responsabilidad que pudiera derivarse de la autoría o falta de: originalidad o contenido de la Tesis presentada de conformidad con el ordenamiento jurídico vigente.

Villahermosa, Tabasco a 07 de Agosto de 2025

Nombre y Firma

LCC. Daniel Domínguez Gómez

Cunduacán, Tabasco a 06 de agosto de 2025
Oficio No. 1322/2025/DACYTI/D

Asunto: Autorización de impresión de Tesis

C. Daniel Domínguez Gómez
Egresado de la Maestría en Tecnologías para el Aprendizaje y el Conocimiento

En virtud de que cumple satisfactoriamente los requisitos establecidos en el Reglamento General de Estudios de Posgrado vigente en la Universidad, informo a Usted que se autoriza la impresión del trabajo recepcional ***"Identificación de factores de deserción universitaria mediante selección de características en modelos de aprendizaje automático"***, para presentar examen y obtener el Grado de Maestro en Tecnologías para el Aprendizaje y el Conocimiento.

Sin otro particular, aprovecho la ocasión para enviarle un afectuoso saludo.

Atentamente

Dr. Óscar Alberto González Gonzalez
Director

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO
DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

C.c.p. Dr. Eddy Arquímedes García Alcocer. - Encargado del Despacho de la Coordinación de Posgrado DACYTI
Archivo.
Consecutivo.

DR. OAGG/EAGA

Carretera Cunduacán-Jalpa Km. 1, Colonia Esmeralda, C.P. 86690
Cunduacán, Tabasco, México
Tel: (993) 358 1500 ext. 6727; (914) 336 0616; Fax: (914) 336 0870
E-mail: direccion.dacyti@ujat.mx

Carta de Cesión de Derechos

Villahermosa, Tabasco a 07 de Agosto.

Por medio de la presente manifestamos haber colaborado como AUTOR(A) y/o AUTORES (RAS) en la producción, creación y/o realización de la obra denominada IDENTIFICACIÓN DE FACTORES DE DESERCIÓN UNIVERSITARIA MEDIANTE SELECCIÓN DE CARACTERÍSTICAS EN MODELOS DE APRENDIZAJE AUTOMÁTICO

Con fundamento en el artículo 83 de la Ley Federal del Derecho de Autor y toda vez que, la creación y/o realización de la obra antes mencionada se realizó bajo la comisión de la Universidad Juárez Autónoma de Tabasco; entendemos y aceptamos el alcance del artículo en mención, de que tenemos el derecho al reconocimiento como autores de la obra, y la Universidad Juárez Autónoma de Tabasco mantendrá en un 100% la titularidad de los derechos patrimoniales por un período de 20 años sobre la obra en la que colaboramos, por lo anterior, cedemos el derecho patrimonial exclusivo en favor de la Universidad.

COLABORADORES

L.C.C. Daniel Domínguez Gómez

Dr. Juan de Dios González Torres

Dr. Arturo Corona Ferreira

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

F7: Respuesta de jurado

Cunduacán, Tabasco, a 18 de marzo de 2025.

MTE. Óscar Alberto González González
Director de la División Académica de Ciencias y Tecnologías de la Información
Presente

En atención a los oficios girados por usted, en los que se nos designa como parte del jurado para efectuar la revisión de la tesis titulada ***"IDENTIFICACIÓN DE FACTORES DE DESERCIÓN UNIVERSITARIA MEDIANTE SELECCIÓN DE CARACTERÍSTICAS EN MODELOS DE APRENDIZAJE AUTOMÁTICO"***, realizada por el ***C. Daniel Domínguez Gómez***, estudiante de la Maestría en Tecnologías para el Aprendizaje y el Conocimiento, nos permitimos informarle que, en virtud de que ha atendido las observaciones realizadas, otorgamos nuestra aprobación para que continúe los trámites para la obtención del grado.

Sin otro particular, aprovechamos la ocasión para enviarle un cordial saludo.

Atentamente integrantes del jurado

Dra. Erika Yunuen Morales Mateos

Dr. Pablo Payró Campos

MTE. Óscar Alberto González González

RECIBIDO
11 ABR. 2025
D.A.C.Y.T.I.
DIRECCIÓN

c.c.p. Dr. Eddy Arquímedes Garcia Alcocer. Encargado del despacho de la Coordinación de Posgrado.
Estudiante.

Dedicatoria

"Porque Jehová da la sabiduría; de su boca salen el conocimiento y la inteligencia."
(Proverbios 2:6)

Quiero dedicar este trabajo a dos pilares fundamentales en mi vida:

A Dios, quien ha sido mi guía constante y fuente de fortaleza durante toda mi vida y en este arduo pero gratificante camino académico. A Él le agradezco por la inteligencia que me ha dado, por cada oportunidad que ha puesto en mi camino y por ser mi roca en los momentos difíciles.

A mis amados padres,Freddy y Leticia a ellos les debo todo lo que soy y todo lo que he logrado hasta el día de hoy. Su apoyo incondicional, su amor infinito y su sacrificio han sido el motor que me impulsó a seguir adelante en cada etapa de mi vida. Nunca podré agradecerles lo suficiente por todo lo que han hecho por mí.

Este trabajo es el resultado de sus enseñanzas, de sus valores inculcados en mí desde pequeño, y de su dedicación incansable para verme triunfar. Cada palabra escrita aquí lleva consigo el amor y gratitud que siento hacia ustedes. ¡Gracias por ser mis padres y por ser mis héroes en cada capítulo de mi vida!, los amo.

Agradecimientos

En primer lugar, agradezco de manera especial a mis directores de tesis, Dr. Juan de Dios González Torres y Dr. Arturo Corona Ferreira, por su guía académica, su paciencia, sus valiosas observaciones y el constante apoyo brindado a lo largo de este proyecto. Su compromiso, experiencia y dedicación han sido fundamentales para mi formación como investigador.

Extiendo mi reconocimiento a mis profesores del Núcleo Académico de MTAC, quienes con su enseñanza, compromiso y motivación, contribuyeron significativamente a ampliar mis conocimientos y fortalecer mis habilidades profesionales.

Agradezco también a mis revisores de tesis, quienes con sus aportaciones, críticas constructivas y sugerencias enriquecieron este trabajo, permitiendo mejorar su calidad y solidez académica.

Tambien deseo expresar mi más sincero y profundo agradecimiento a todas las personas, compañerpos y amigos que siempre me apoyaron de alguna u otra manera, muchas gracias.

Finalmente, mi más profundo agradecimiento al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el financiamiento otorgado, el cual hizo posible mi dedicación plena a los estudios y la culminación de esta etapa académica.
A todos ustedes, mi gratitud y reconocimiento por formar parte de este logro.

Índice de Contenido

Índice de tablas

Índice de figuras

Título

Identificación de factores de deserción universitaria mediante selección de características en modelos de aprendizaje automático

Resumen

El abandono escolar supone una grave amenaza para el desarrollo personal y social, ya que limita las oportunidades laborales y perpetúa las desigualdades económicas. Este estudio aborda el reto de identificar a los estudiantes que abandonan los estudios centrándose en la selección de características o variables relevantes. Aplicamos cinco técnicas de selección de características: ANOVA, información mutua, selección secuencial hacia delante, eliminación recursiva de características y el operador de selección y reducción mínima absoluta, y destacamos las características sociodemográficas clave. Los resultados mostraron que el modelo KNN de información mutua obtuvo la precisión y la puntuación F1 más altas (99,05 %), mientras que SFS proporcionó resultados óptimos para los árboles de decisión y las máquinas de vectores de apoyo. Estos resultados ponen de manifiesto el papel fundamental de las técnicas de selección de características a la hora de identificar las variables que afectan significativamente a la eficacia de los modelos predictivos.

Palabras claves: abandono escolar, predicción, educación, técnicas de selección de características, aprendizaje automático, algoritmos.

Title

Identifying key factors of university dropout using feature selection in machine learning models

Abstract

School dropout poses a significant threat to personal and social development, limiting employment opportunities and perpetuating economic inequality. This study addresses the challenge of identifying student dropouts by focusing on selecting relevant features or variables. We applied five feature selection techniques: ANOVA, mutual information, sequential forward selection, recursive feature elimination, and the least absolute shrinkage and selection operator, highlighting key sociodemographic features. Fifteen machine learning models, including Decision Trees, Support Vector Machines, and K-nearest neighbors, were trained using these techniques and evaluated through cross-validation. The results showed that the mutual information KNN model achieved the highest precision and F1 score (99.05%), while SFS provided optimal results for Decision Trees and Support Vector Machines. These results highlight the critical role of feature selection techniques in identifying the variables that significantly affect the effectiveness of predictive models.

Keywords: dropout, prediction, education, feature selection techniques, machine learning, algorithm

Capítulo I. Introducción

1.1. Planteamiento del problema

La deserción escolar es un problema social en México, una de las problemáticas vigentes en el sistema educativo, pues, cuando se han implementado medidas para combatir el fenómeno, estas son insuficientes, ya que esto se sigue presentando en cada ciclo escolar y dentro de los tres niveles de educación (Pulso,2018). Debido a la situación económica que afecta a nivel nacional, se ha observado un aumento en la deserción escolar en la Universidad Juárez Autónoma de Tabasco, específicamente entre el uno y dos por ciento de la población estudiantil, este incremento se le puede atribuir a la falta de recursos para continuar con los estudios. Con la esperanza de abordar este problema, se busca destinar mayores recursos económicos, especialmente en caso de un posible aumento en el presupuesto. El objetivo es asignar fondos adicionales para ampliar la cantidad de becas disponibles, con la intención de mitigar, en la medida de lo posible, las dificultades financieras que enfrentan los estudiantes y así prevenir la deserción escolar (Ruiz, 2013).

Para comprender el fenómeno de la deserción escolar en el ámbito universitario, es crucial identificar todos los factores que influyen en la decisión de abandonar los estudios. De acuerdo con investigaciones recientes, se ha establecido que la minería de datos educativos (MDE o Educational Data Mining) y el aprendizaje automático (machine learning) son herramientas relevantes para analizar situaciones en este contexto (Clow, 2013). En los últimos años, la deserción escolar y los elevados índices de reprobación son las problemáticas más importantes en instituciones educativas, provocando bajos índices de eficiencia terminal (Martínez *et al.* 2013; Timarán *et al.* 2013). Basados en los reportes del año 2018 al 2020, se puede evidenciar el incremento en la deserción de estudiantes; en este sentido, el Banco Mundial el 2018, informó que aproximadamente el 30% de los estudiantes universitarios interrumpieron sus estudios debido a distintas

causas, además, según publicación en el diario El Comercio (Alayo, 2020), unos 174,000 (18,6%) estudiantes universitarios dejarían sus estudios en el año 2020.

Según Delen (2010), la retención universitaria se presenta como una variable de gran importancia en las instituciones de educación superior. Esto se debe a que tiene un impacto directo en el prestigio y la reputación de la marca de la organización educativa. Por lo tanto, resulta crucial monitorear y controlar los índices de deserción, ya que este factor a menudo es considerado en la clasificación de las universidades en diversos rankings. Asimismo, Dénle sugiere que la gestión de las matrículas universitarias debería tener en cuenta esta variable para un manejo más efectivo y estratégico.

En el estado de Tabasco, a pesar de los notables avances en la reducción de los índices de deserción escolar, persisten desafíos significativos. La carencia de recursos económicos, los problemas emocionales, las adicciones, embarazos no deseados y la elección inadecuada de carreras continúan siendo los principales motivos que llevan a los jóvenes a abandonar sus estudios en los niveles medio y superior en la región, lamentablemente, este tema sigue generando preocupaciones y provocando reacciones tanto a nivel federal como estatal. Ver Tabla 1

Tabla 1

Índices de deserción escolar universitaria, nivel nacional, estatal y en la UJAT

Año	México (Nacional)	Tabasco	UJAT
2020	8.5	9.9	30.0
2021	8.8	7.9	12.0
2022	8.1	9.1	6.1
2023	7.2	8.0	4.0

Nota: Elaboración propia con base con información de (INEGI 2024, SEP 2024 Y UJAT 2024).

Aunque se han establecido estrategias para reducir la tasa de deserción escolar, se han implementado programas como “Apadrina un estudiante” para los jóvenes de bajos recursos, más seguimiento con tutorías de profesores para orientar a los estudiantes, y con psicólogos en cada División para tratar los temas emocionales, etc. (Rodríguez, R. I. G., & Mazariego, C. R. R., 2020).

En las instituciones universitarias, se deberían manejar datos detallados sobre el perfil de los estudiantes desde el momento de su matriculación, y disponer de esta información relevante en diversos aspectos. A lo largo de los semestres académicos, se continúa registrando datos adicionales de importancia, que abarcan desde información económica hasta datos académicos y de salud (Valero Cajahuanca *et al.* 2022).

En el ámbito educativo, se han aplicado técnicas de ML para diversas finalidades, incluyendo la predicción de la deserción estudiantil y el apoyo al rendimiento académico (Cruz, E., González, M., & Rangel, J., 2022).

El problema al que se enfrentan muchas universidades en México radica en la falta de datos y variables específicas que faciliten el desarrollo de un modelo basado en técnicas de selección de características y aprendizaje automático (ML) que puedan ayudar a reducir los índices de deserción escolar.

Esta investigación se enfocará en cubrir estas deficiencias, con el objetivo de desarrollar un modelo predictivo fiable que permita a las instituciones educativas identificar oportunamente a los estudiantes en riesgo de deserción.

1.2. Pregunta de investigación

¿Cuál es la precisión de los modelos de aprendizaje automático utilizando selección de características en la identificación de la deserción escolar universitaria?

¿Cuáles son las variables más relevantes que pueden ayudar a predecir la deserción escolar universitaria después de aplicar técnicas de selección de características en modelos de aprendizaje automático?

1.3. Hipótesis o supuesto

Los algoritmos de aprendizaje automático pueden ayudar a predecir de manera efectiva la deserción escolar universitaria mediante la identificación de variables clave, tales como los factores demográficos, socioeconómicos, académicos y conductuales, mejorando la precisión del modelo a través de la selección de las características más relevantes.

1.4. Objetivos

1.4.1. Objetivo general

Optimizar la identificación de los factores claves que influyen en el abandono escolar mediante el uso de algoritmos de aprendizaje automático y técnicas avanzadas de selección de características.

1.4.2. Objetivos específicos

- Entrenar y comparar diferentes algoritmos de clasificación de aprendizaje automático para determinar cuáles ofrecen el mejor rendimiento
- Identificar las variables que influyen de manera directa en la eficacia de los algoritmos de aprendizaje automático, empleando técnicas de selección de características.

1.5. Justificación

En los últimos años, se ha tenido un avance tecnológico sin precedentes en diversas áreas, con especial relevancia en el ámbito educativo, según lo destacado por Sánchez (2009). La literatura respalda la idea de que el uso de herramientas tecnológicas, cuando

se implementa con prácticas adecuadas, puede potenciar los procesos de enseñanza y aprendizaje, como señalan Torres y Cobo (2017).

Uno de los propósitos del análisis de datos educativos es identificar patrones y realizar predicciones que ayuden a describir el progreso académico de los estudiantes. Sin embargo, es esencial recopilar datos sobre las características de los estudiantes, considerando el contexto, para lograr una comprensión más completa de los resultados obtenidos. Entre estas características se incluyen factores socioeconómicos, así como información sobre la familia y el historial escolar del estudiante. (Rico Páez & Gaytán Ramírez ,2022).

En el análisis de datos se utilizan diversas metodologías, técnicas y algoritmos, como señala Peña (2014). La anticipación del rendimiento académico se realiza con varios propósitos, incluyendo la identificación del riesgo de abandono o la probabilidad de deserción por parte de los estudiantes.

Es conocido que una de las principales inquietudes de las Instituciones de Educación Superior ha sido mejorar sus indicadores de eficiencia terminal y logro educativo, al mismo tiempo que reducir el rezago y la deserción escolar.

Sin embargo, según los datos proporcionados por Díaz de Cosió (1998), a nivel nacional, en promedio, de cada 100 alumnos que inician una carrera de nivel licenciatura, solo 60 completan sus materias en un plazo de cinco años y únicamente 20 de estos logran obtener el título, lo que se traduce en una eficiencia de titulación del 20%. La deserción escolar representa un problema educativo que impacta tanto el desarrollo individual del estudiante que deja de asistir a la escuela como la sociedad en la que se desenvuelve (Martínez, 2013).

Es por eso por lo que la deserción temprana de estudiantes representa un desafío significativo para las universidades en la actualidad. Diversas técnicas de minería de datos y aprendizaje automático se han empleado para identificar a aquellos estudiantes

en riesgo de abandonar sus estudios. Estos modelos, que utilizan datos sociodemográficos y calificaciones del nivel anterior, han demostrado una precisión lo suficientemente sólida como para respaldar la implementación de programas de retención. La inclusión de calificaciones de los primeros semestres mejora aún más la exactitud de estos modelos predictivos.

A pesar de los avances, los modelos existentes pueden incurrir en errores de clasificación, lo que tiene consecuencias notables, los estudiantes no detectados como en riesgo pueden quedar excluidos de los programas de retención, mientras que aquellos que no enfrentan un riesgo real pueden consumir recursos adicionales, este fenómeno destaca la necesidad de afinar y perfeccionar los modelos predictivos para minimizar los errores de clasificación y garantizar una asignación eficiente de recursos en los esfuerzos de retención estudiantil.

En los últimos quince años ha surgido la Minería de Datos Educativos (EDM) como una nueva área de aplicación que se ocupa de desarrollar, investigar y aplicar métodos informáticos para detectar patrones en grandes colecciones de datos educativos que, de otro modo, serían difíciles o imposibles de analizar debido al enorme volumen de datos en el que se encuentran (Baker y Yacef, 2009), (Romero y Ventura, 2013). Una de las aplicaciones más antiguas y conocidas de la GED es la predicción del rendimiento de los alumnos, en la que el objetivo es estimar el valor desconocido del rendimiento, los conocimientos, la puntuación o la nota de un alumno (Romero y Ventura, 2007), (Romero y Ventura, 2010), (Wolff et al., 2014), (Yo y Kim, 2014). La clasificación es la técnica más empleada para resolver este problema descubriendo modelos predictivos del rendimiento de los alumnos basados en datos históricos de los mismos (Hämäläine y Vinni, 2011), (Vialardi *et al.* 2011), (Romero *et al.* 2013). Sin embargo, para la predicción temprana del abandono de los estudiantes se convierte en una tarea más difícil porque la tarea de clasificación tradicional no se las arregla bien con la naturaleza temporal de este tipo específico de datos, ya que normalmente considera que todos los atributos están siempre disponibles (Antunes, 2010).

Como se ha señalado anteriormente, el abandono escolar conlleva consecuencias serias para los individuos, las instituciones educativas y la sociedad en su conjunto, según investigaciones como las de Liem *et al.* (2001) y Latif *et al.* (2015). En la actualidad, nos encontramos en una etapa en la que las razones detrás del abandono son diversas y complejas, pero se comprenden hasta cierto punto, ya que factores como el estatus sociodemográfico, la salud mental y el entorno universitario constituyen la base del fenómeno del abandono, según estudios realizados por Heublein (2013) y Larsen *et al.* (2013).

Consideramos que la prioridad radica en la identificación de estos estudiantes, porque, aunque implementemos programas de asesoramiento y tutorías de alta calidad para brindar apoyo a estos alumnos, carece de sentido si no los detectamos antes de que abandonen la institución.

~~Por lo tanto,~~ cualquier mejora, aunque sea relativamente pequeña, puede tener un impacto considerable tanto en la institución como en el estudiante. Este impacto es particularmente crucial para los individuos, ya que no es exagerado afirmar que el abandono escolar podría tener consecuencias significativas para su futuro, según las investigaciones de Liem *et al.* (2001).

La identificación temprana de los estudiantes vulnerables que son propensos a abandonar sus cursos es crucial para el éxito de cualquier estrategia de retención escolar. Y, para tratar de reducir el problema mencionado, es necesario detectar lo antes posible a los estudiantes en situación de riesgo y, de este modo, prestarles cierta atención para evitar que abandonen sus estudios e intervenir precozmente para facilitar la retención de los estudiantes (Heppen y Bowles, 2008).

1.6. Metodología

1.6.1. Enfoque de Investigación

El enfoque metodológico de la investigación fue cuantitativo, ya que al utilizar técnicas de minería de datos se analizan grandes cantidades de datos, se pueden hallar patrones y modelos en investigación educativa, de acuerdo con Dicovskiy y Pedroza (2018) las bases de datos académicas son un material importante en cualquier investigación educativa, y deberían ser estudiadas por minería de datos, como un método innovador dentro los métodos tradicionales de investigación cuantitativa.

La investigación que adopta el enfoque cuantitativo se caracteriza por abordar fenómenos que son susceptibles de medición, es decir, aquellos que pueden asignarse a un número concreto, como, por ejemplo: cantidad de hijos, edad, peso, estatura, aceleración, masa, nivel de hemoglobina, cociente intelectual, entre otros.

Este enfoque se vale de técnicas estadísticas para el análisis de los datos recolectados y su objetivo principal consiste en describir, explicar, prever y controlar de manera objetiva las causas de un fenómeno, así como prever su ocurrencia a partir de la revelación de estas causas, sus conclusiones se fundamentan en el uso riguroso de la medición o cuantificación, tanto en la recopilación de resultados como en su procesamiento, análisis e interpretación, siguiendo el método hipotético-deductivo. En este sentido, su aplicación encuentra un terreno más extenso en las ciencias naturales, como biología, química, física, neurología, fisiología, psicología, entre otras. (Kerlinger, 2002).

De igual manera tiene un alcance descriptivo, puesto que explica el comportamiento de una variable. Según Hernández, Fernández y Baptista (2014), la investigación cuasi experimental mantiene la rigurosidad científica, el tipo de estudio también es transaccional descriptivo.

1.7. Población de Estudio

Población, es el conjunto de personas u objetos de los que se desea conocer algo en una investigación. "El universo o población puede estar constituido por personas, animales, registros médicos, los nacimientos, las muestras de laboratorio, los accidentes viales entre otros". (Pineda *et al.* 1994). Para esta investigación nuestra población de estudio se centra en conjunto llamado *"Predict students' Dropout and Academic Success"* que está disponible públicamente en el sitio web *Kaggle*. Este conjunto de datos contiene datos académicos, macroeconómicos y socioeconómicos, junto con información sobre la matrícula de los estudiantes de primer y segundo semestres, con variables académicas a fines de nuestra investigación, tales como como son nombre, genero, avance curricular, promedio general etc.

1.8. Instrumento para la recolección de datos

Como menciona (Hernández *et al.* 2010), el propósito del instrumento de recolección de datos es establecer las condiciones necesarias para la medición. Los datos representan conceptos que abstraen el mundo real, capturando elementos sensoriales que pueden ser percibidos de manera directa o indirecta. Todo lo empírico puede ser medido.

Toda medición o instrumento de recolección de datos debe reunir dos requisitos esenciales; confiabilidad y validez. La confiabilidad de un instrumento de medición se refiere al grado de precisión o exactitud de la medida, en el sentido de que si aplicamos repetidamente el instrumento al mismo sujeto u objeto produce iguales resultados. Es el caso de una balanza o de un termómetro, los cuales serán confiables si al pesarnos o medirnos la temperatura en dos ocasiones seguidas, obtenemos los mismos datos. (Hernández, 2010)

la recolección de datos fue a través de fuente secundaria, el análisis secundario de documentos o de datos documentales se refiere a la evaluación adicional de un conjunto de datos primarios, con el objetivo de proporcionar nuevas interpretaciones y conclusiones, o presentar los hallazgos de manera distinta a la original. (Sierra Bravo, 2003).

De acuerdo con Sierra Bravo (2003), el análisis secundario de datos requiere cumplir con dos condiciones fundamentales: primero, realizar un nuevo análisis o reanálisis de datos que ya fueron obtenidos y evaluados anteriormente; segundo, que este nuevo análisis no repita los análisis previamente realizados con esos datos, sino que aporte tratamientos diferentes y nuevas interpretaciones adicionales, además, es crucial que en todos los casos se enfatice la importancia de la validez y fiabilidad de la información, considerando aspectos teóricos, epistemológicos y metodológicos, especialmente en relación con el proceso de recopilación de datos.

De acuerdo con Selltiz (1980), los datos secundarios son útiles para generar hipótesis de investigación, pero su aprovechamiento requiere la capacidad de plantear una amplia variedad de preguntas relacionadas con el problema de investigación, por lo tanto, el principio fundamental para utilizar estadísticas disponibles radica en ser flexible en la formulación de las preguntas de investigación (Selltiz, 1980, p. 510).

Capítulo II. Marco teórico

2.1.1. Deserción en instituciones de educación superior

Para empezar, es importante definir claramente lo que se entiende por deserción estudiantil en las instituciones de educación superior. Aunque no existe una definición universalmente aceptada, hay elementos comunes que permiten identificar este fenómeno. La deserción puede ser analizada desde el comportamiento individual de los estudiantes y cómo estos influyen en la decisión de abandonar sus estudios, hasta una visión más amplia que considera factores colectivos, institucionales y a nivel nacional. Para comprender mejor este tema, es fundamental revisar las medidas que los gobiernos han implementado para reducir la deserción escolar y los factores que históricamente han influido en su aparición. (Camargo García, 2020).

2.1.2. Deserción según la perspectiva estatal o nacional

En (Vicent Tinto, 1989) se indica que el concepto es diferente cuando la perspectiva es estatal. Por ejemplo, la dimisión entre instituciones de educación públicas no puede representar deserciones en el sentido estricto de la palabra, ya que se trata de cambios internos realizados en el sector estatal. No obstante, si hay una salida de estudiantes con destino a instituciones privadas, es probable que las dimisiones o abandonos se tengan en cuenta como deserciones. En este sentido, solo se considerarán deserciones las personas que abandonan todo el sistema de educación superior.

2.1.3. Deserción desde el punto de vista Institucional

Según Vicent Tinto (1989), definir la deserción mediante un enfoque institucional es, en ciertos aspectos, un trabajo más sencillo que hacerlo desde el punto de vista individual. En otros aspectos, aunque los hay, es ampliamente más difícil. El autor plantea que es más sencillo porque todas las personas que abandonan una institución de educación superior logran valorar las razones expuestas para ser catalogados como desertores. Los

estudiantes que abandonan crean una vacante que otros podrían haber ocupado. En consecuencia, el autor argumenta que esto genera problemas financieros al generar un desequilibrio en los ingresos de las instituciones. Este problema es más evidente en el sector privado, en el que las matrículas suponen una parte fundamental de los ingresos, aunque también afecta al sector público debido a la falta de presupuesto.

2.2. Marco referencial

2.2.1. Modelos predictivos usando algoritmos de aprendizaje automático para la deserción escolar

En la actualidad, gracias al desarrollo tecnológico, se han realizado muchos estudios en el ámbito de la ciencia de datos y el aprendizaje automático. Esta disciplina se ha generalizado en áreas como la predicción del riesgo de fracaso de los estudiantes, la predicción de las calificaciones de los exámenes finales y la identificación temprana de los estudiantes fracasados. (R. Z. Pek *et al.* 2021)

En este estudio, Talamas y Ceballos, (2023) proponen una técnica de *ensemble stacking* como medio para combinar modelos predictivos que, con un número limitado de variables, sean capaces de alcanzar los resultados deseados para la predicción del abandono escolar temprano. Los resultados indican que la aplicación de esta técnica en un programa de intervención sería rentable. Por el contrario, Cuevas-Chávez *et al.* (2023), se centraron más en la optimización de parámetros y técnicas de Re-muestreo (*Adaptive Synthetic, SMOTE-SVM, y SMOTE+ENN*) para abordar el desequilibrio de datos, evaluando el rendimiento de clasificadores como *Random Forest, Support Vector Machine, y XGBoost*. La combinación de *SMOTE+ENN* con el clasificador SVM mostró el rendimiento óptimo, con una precisión del 94,11% en los modelos predictivos.

Por otra parte, estudios hechos por Dharmawan *et al.* (2018) tomaron datos de los estudiantes que estudian en varias universidades de Indonesia utilizando un muestreo aleatorio simple. Tomando información demográfica, motivación, financiera, interacción

social y personalidad, analizaron con los clasificadores SVM, arboles de decisión y K-NN y obtener una precisión de la deserción del 66% para los árboles de decisiones y SVM.

En investigaciones de Segura-Morales & Loza-Aguirre, (2018) buscaron determinar cómo los factores socioeconómicos afectan los logros educativos de los estudiantes de secundaria, se consideraron datos socioeconómicos y académicos correspondientes a más de diez años de registros obtenidos de la principal universidad de un país de la región andina. Utilizando algoritmos de clasificación y técnicas de aprendizaje automático para determinar qué factores influyen más en el rendimiento académico. Se encontró que las becas académicas, la edad, el condado y el grado de la escuela secundaria influyen en el rendimiento académico de los estudiantes.

De acuerdo con los análisis de Pérez *et al.* (2019), hacen una comparación de indicadores de desempeño del modelo actual de deserción de la Universidad del Bío-Bío (UBB) en Chile, que se basa en la técnica de regresión logística y se compara con un nuevo modelo basado en árboles de decisión. La comparación muestra que la predicción de la deserción escolar del modelo propuesto obtiene una exactitud del 86%, una precisión del 97% con una tasa de error del 14%.

En el siguiente estudio de Lee & Chung, (2019) el objetivo fue mejorar el funcionamiento de un sistema de alerta temprana de deserción escolar, abordando el problema del desequilibrio de clase utilizando las técnicas de sobre muestreó de minorías sintéticas *(SMOTE)* y los métodos de conjunto en el aprendizaje automático. Evaluando los clasificadores capacitados con curvas tanto de características operativas del receptor (ROC) como de *precision–recall (PR).* Utilizando las muestras de grandes datos de los 165.715 estudiantes de secundaria del Sistema Nacional de Información Educativa (NEIS) de Corea del Sur. Se entrenaron con cuatro algoritmos de clasificación: bosque aleatorio (RF), árbol de decisión impulsado (BDT), bosque aleatorio con *SMOTE (SMOTE+ RF),* y árbol de decisión impulsado con *SMOTE (SMOTE + BDT*). El árbol de decisión impulsado mostró el mejor desempeño.

Algunos de los modelos más comunes incluyen *Random Forest*, Redes Neuronales, *Support Vector Machines* y Regresión Logística. Estos modelos han sido objeto de comparaciones en estudios como el realizado por un equipo del "Instituto Tecnológico de Costa Rica" (Solís *et al.* 2018) y en investigaciones realizadas en el Instituto Tecnológico de Karlsruhe, donde se emplearon árboles de decisión y regresión logística (Kemper *et al.* 2020).

El trabajo de Berens *et al.* (2019), implementó un sistema de detección temprana para identificar el abandono escolar. Este estudio utilizó modelos como la regresión logística, Random Forest con *bagging* y el algoritmo *AdaBoost.*

Otros trabajos que han utilizado técnicas de aprendizaje automático en la creación de modelos predictivos: Salal *et al*. (2019), los cuales construyeron modelos de predicción para predecir el rendimiento académico de estudiantes con algoritmos como Naïve Bayes, árbol de decisión, entre otros. Este estudio evidencia cómo ciertas características ejercen influencia en el rendimiento de los estudiantes. Contreras *et al.* (2020) emplearon modelos basados en algoritmos de aprendizaje automático para determinar qué factores influyen en la interrupción o continuación de los estudios de los alumnos de ingeniería. Por su parte, Castrillón *et al.* (2020) llevaron a cabo una investigación en la que utilizaron técnicas de árboles de decisión para predecir el rendimiento académico de estudiantes de nivel superior. En investigaciones más recientes, los factores vinculados al rendimiento académico de estudiantes universitarios han sido examinados mediante el uso de regresiones lineales (Gutiérrez *et al.* 2021).

2.2.2. Técnicas de selección de características aplicadas en el desarrollo de modelos predictivos para la deserción escolar

Existe una etapa inicial de preprocesamiento a la hora de implementar modelos computacionales, ya que los datos requieren un tratamiento adecuado que permita conocer su comportamiento y su influencia en la selección de técnicas, el entrenamiento y los resultados (Kuz y Morales, 2023),esta etapa incluye la limpieza (datos incorrectos, atípicos o vacíos), la integración, la transformación (discretización, normalización), la

selección de características, la reducción de características y el equilibrio de clases. En algunos estudios no se menciona en detalle, pero existe una necesidad de fiabilidad de la información que está mediada por la calidad de los datos (Pradeep *et al*. 2019), se han identificado variaciones en los resultados finales de los métodos utilizados para la predicción, que se han asociado a deficiencias en el preprocesamiento.

Eckert y Suénaga (2015), afirman que el preprocesamiento se considera una etapa extensa y, al mismo tiempo, fundamental, porque de él dependen los resultados posteriores obtenidos en el entrenamiento y la evaluación de los algoritmos, en otras palabras, es una etapa crucial antes de implementar cualquier técnica de data mining educativo. Márquez-Vera *et al*. (2013) mencionan dos problemas a abordar en la minería de datos educativos:

(i) La dimensión del número de características

(ii) El desequilibrio de clases. Por otra parte, un estudio de Delen *et al.* (2020) demostró que la reducción de la dimensionalidad de los datos y el uso de datos sintéticos (para clases desequilibradas) favorecían un aumento del 6% en la precisión de los algoritmos de aprendizaje automático. Por todo lo anterior, la calidad de los datos influye significativamente en los resultados de los algoritmos de aprendizaje automático.

Tabla 2.1

Trabajos relacionados que aplicaron diversas técnicas de selección de características.

Trabajo	Algoritmos de aprendizaje automático	Descripción del conjunto de datos	Técnica de selección de características	Resultados
Youssef, M *et al.,* (2019)	• SVM, • Decision Tree • Naive Bayes	El conjunto de datos es anónimo y fue extraído de la plataforma OpenEdx. Contiene15 características y 5327 registros de estudiantes matriculados en dos periodos diferentes	Sequential Forward Selection, Sequential Backward Selection,	After SVM AUC 0.972 Precision 0.988 Recall 0.980 F1-score 0.984

	• LR • KNN		Recursive Feature Elimination	Before RF SFS AUC 0.962 Precision 0.989 Recall 0.964 F1-score 0.976
Zapata-Medina *et al.,* (2024)	• SVM • RF • GBT	El conjunto de datos de una universidad pública no mencionada contiene información personal, socioeconómica y académica desde 2016 hasta 2019, con 1,865 registros y 29 características (16 demográficas, 12 académicas), además de la etiqueta de clase.	FSmRMR, FS Boruta FS LASSO, FS GA, FS PSO	After RF Precision 0.920 Recall 0.570 F1-score 0.700 Before FS LASSO GBT Precision 0.924 Recall 0.632 F1-score 0.751
Li *et al.,* (2018)	• LR • SVM • RF	Conjunto de datos de la plataforma MOOC XuetangX para la KDD CUP 2015. El conjunto de datos incluye 39 cursos, 79,186 estudiantes y 120,542 inscripciones, con 8,157,277 registros de entrenamiento con etiquetas completas.	Mutual information (MI), random forest (RF), and recursive feature elimination (RFE)	After LR AUC 0.8630 Precision 0.8567 Recall 0.8629 F1-score 0.8449 Before LR RF AUC 0.8629 Precision 0.8573 Recall 0.8633 F1-score 0.8472

Nota: Elaboración propia con base con a la información de Youssef, M *et al.* (2019), Zapata-Medina *et al.* (2024), Li *et al.* (2018).

Estas Investigaciones sobre la selección de características para la creación de modelos de predicción del abandono escolar han demostrado mejoras significativas en la precisión de los modelos de aprendizaje automático.

Estudios como los realizados por Youssef *et al.* (2019), Zapata-Medina *et al.* (2024) y Li *et al.* (2018) han empleado técnicas como la selección secuencial de características (SFS), *Boruta y LASSO*, que han dado como resultado mejoras significativas en métricas clave como AUC, precisión, recuperación y puntuación F1 en modelos que incluyen la

máquina de vectores de soporte (SVM), el bosque aleatorio (RF) y los árboles de aumento gradual (GBT). Estos resultados ponen de relieve la importancia de una selección adecuada de características para optimizar el rendimiento predictivo y mejorar la capacidad de los modelos para identificar a los estudiantes en riesgo con mayor precisión.

2.2.3. Modelos de clasificación de aprendizaje automático

Son algoritmos de inteligencia artificial que, mediante procesos de aprendizaje supervisado y no supervisado, posibilitan el entrenamiento de una máquina. Estos algoritmos generan modelos matemáticos con la capacidad de generalizar comportamientos.

Existen diversos modelos de aprendizaje automático, entre ellos, la Regresión logística, Árboles de decisión y *K-Means*. La Regresión logística, por ejemplo, se destaca como una herramienta versátil para llevar a cabo clasificaciones de múltiples clases. En términos visuales, al representar la regresión se obtiene una curva con forma de S, lo cual facilita la división de los datos en grupos distintos. (Microsoft Azure, 2022).

2.2.4. Construcción de modelos predictivos

Como menciona Rico Páez & Gaytán Ramírez (2022). Las técnicas de aprendizaje automático posibilitan la creación de modelos especializados utilizando un conjunto de registros para un resultado específico. En términos sencillos, construir este tipo de modelo requiere contar con información detallada, conocida como "datos de entrenamiento", y aplicar una técnica de aprendizaje automático. Estas técnicas incorporan varios mecanismos para elaborar modelos predictivos, los cuales pueden manifestarse como algoritmos, ecuaciones, estructuras u otras formas relevantes.

La construcción de modelos predictivos es un proceso sistemático que va desde la definición del problema hasta la implementación del modelo en un entorno real. A continuación, se presenta una tabla con trabajos relevantes que aplican aprendizaje automático para la predicción de la deserción escolar. Ver tabla 3

Tabla 3

Trabajos relacionados con la construcción de modelos predictivos aplicados a la deserción escolar

Núm.	Año	Título del Artículo	Objetivo del articulo	Resultados	Conclusiones
1	2024	School Dropout Prediction with Class Balancing and Hyperparameter Configuration	predecir las tasas de deserción escolar mediante el uso de técnicas de balanceo de clases y configuración de hiperparámetros. Esto busca mejorar la precisión de las predicciones de deserción	La predicción de deserción escolar mostró que la metodología utilizada, que incluye técnicas de balanceo de clases y configuración de hiperparámetros, mejoró significativamente la precisión de los modelos predictivos.	la implementación de técnicas de balanceo de clases y la adecuada configuración de hiperparámetros resultaron en una mejora significativa en la precisión de los modelos de predicción de deserción escolar.
2	2023	Machine Learning Model for Student Drop-Out Prediction Based on Student Engagement	El propósito del artículo es demostrar cómo se pueden utilizar las técnicas de aprendizaje automático para predecir la deserción estudiantil en función de la participación y el rendimiento de los estudiantes.	Los resultados indican que diversos tipos de participación estudiantil juegan un papel crucial en la predicción de la deserción escolar y los logros finales de los puntos ECTS.	El trabajo demuestra que la deserción estudiantil puede predecirse utilizando diferentes técnicas de ML, como el aprendizaje supervisado y el aprendizaje no supervisado.

3	2023	A stacking ensemble machine learning method for early identification of students at risk of dropout.	Proponer el uso de un método de aprendizaje automático por conjuntos (stacking ensemble) para la detección precoz de estudiantes en riesgo de abandono mediante un enfoque de clasificación El método ensemble combina los resultados de varios modelos diferentes para obtener un modelo combinado mejorado	Un apilamiento de los algoritmos, con los cuales se buscó cuáles eran los más precisos de acuerdo con las variables dadas y predecir de manera eficaz los alumnos más propensos a desertar	Este trabajo aporta una serie de modelos de detección precoz que pueden utilizarse tanto para asignar programas de regularización a alumnos que podrían estar matriculándose con puntuaciones inferiores a las ideales, como para posibles intervenciones en alumnos con dificultades.
4	2022	Using Machine Learning Techniques to Predict Learner Drop-out Rate in Higher Educational Institutions	El propósito del artículo es abordar la preocupación de las altas tasas de deserción entre los alumnos de las instituciones de educación superior y desarrollar un algoritmo de aprendizaje automático que pueda predecir con precisión las tasas de deserción estudiantil e identificar los factores que contribuyen a la deserción y retención.	El algoritmo de bosque aleatorio (RF) surgió como el algoritmo de mejor rendimiento, con una precisión de 70.98% y 69.74% para las implementaciones de validación cruzada de 10 y 5 veces, respectivamente	La implementación del modelo basado en RF para la predicción futura puede ayudar a identificar a los alumnos con pensamientos de deserción para recibir asesoramiento académico inmediato.

5	2023	Machine Learning Prediction of University Student Dropout: Does Preference Play a Key Rôle?	El artículo tiene como objetivo desarrollar métodos estadísticos para la detección temprana de la deserción estudiantil y atender los impactos económicos y sociales de este fenómeno.	El estudio encontró que la detección de deserción mejoró después de los resultados del primer semestre y que el rendimiento académico siempre fue una variable relevante en la predicción de la deserción estudiantil.	Los modelos de aprendizaje automático, como las máquinas vectoriales de soporte, los árboles de decisión y las redes neuronales artificiales, arrojaron los mejores resultados en la predicción de la deserción estudiantil

Nota: Elaboración propia con base con a la información de los artículos citados en la tabla

2.3. Marco conceptual

2.3.1. Deserción escolar

El debate sobre la deserción comienza con su propia definición, según Tinto (1989), el estudio de la deserción en la educación superior y media es sumamente complejo, ya que implica no solo una diversidad de perspectivas, sino también una amplia gama de distintos tipos de abandono, adicionalmente, afirma que ninguna definición puede abarcar por completo la complejidad de este fenómeno, quedando a discreción de los investigadores la elección de la definición que mejor se adecue a sus objetivos y al problema que están investigando. De igual manera, Tinto (1982) conceptualiza la deserción como la situación que experimenta un estudiante cuando aspira, pero no logra completar su proyecto educativo. Indica que se puede clasificar como desertor a aquel individuo que, siendo estudiante, no participa en actividades académicas durante tres semestres consecutivos.

En algunas investigaciones, este comportamiento se denomina "primera deserción" (*first drop out*), ya que resulta difícil determinar si, transcurrido este periodo, el estudiante

retomará o no sus estudios, o si optará por iniciar otro programa académico en otra institución (Tinto, 1989; Cabrera *et al.* 1992 y 1993; Adelman, 1999).

González (2005) propone diferenciar dos tipos de abandonos en los estudiantes universitarios, considerando tanto el factor temporal (inicial, temprano y tardío) como el aspecto espacial (institucional, interno y del sistema educativo).

Basándonos en esta definición, al referirnos a estudiantes universitarios, podemos distinguir dos tipos de abandono según se clasifiquen en función de criterios relacionados con el tiempo y el espacio. En cuanto al tiempo, la deserción puede subdividirse en:

➢ Deserción precoz: El individuo que, habiendo sido aceptado por la universidad, no se matricula

➢ Deserción temprana: Aquel que abandona sus estudios en los primeros semestres de la carrera.

➢ Deserción tardía: Quien abandona los estudios en los últimos semestres, es decir, una vez cursados al menos la mitad de los semestres establecidos en el programa académico. (Castaño, Gallón, Gómez, y Vásquez, 2008)

2.3.2. Factores de la deserción

La literatura especializada señala que hay diversas razones comunes por las cuales los estudiantes universitarios abandonan sus estudios. Estas incluyen problemas vocacionales, situación económica y rendimiento académico (Centro Microdatos, 2008); asuntos laborales, de salud, maternidad, indisciplina o insatisfacción académica (Gracia, 2015); y rasgos de personalidad vinculados a factores psicológicos, sociológicos, económicos, organizacionales y de adaptación e integración (Díaz y Tejedor, 2017).

Las investigaciones sobre las causas de la deserción son de naturaleza multifactorial, razón por la cual es fundamental comprender y esclarecer cada uno de estos factores para realizar un análisis más completo (Liz, 2011).

2.3.3. Problemas financieros

Un considerable número de estudiantes opta por abandonar la universidad debido a dificultades económicas. El elevado costo de la matrícula, los gastos asociados con libros y materiales de estudio, así como la necesidad de trabajar para sostenerse o respaldar a sus familias, constituyen obstáculos significativos que complican la continuidad de sus estudios universitarios. Las becas otorgadas a los estudiantes juegan un papel determinante en su capacidad para mantenerse en la universidad. Se ha notado que las tasas de deserción varían según la cantidad y duración de la ayuda financiera que reciben los estudiantes (Ishitani & desJardins, 2002), y se ha observado que incluso enfrentar apuros económicos provoca un impacto significativo en el abandono temprano (Ozga & Sukhmandan, 1998).

De acuerdo con Rodríguez Lagunas y Hernández Méndez (2008), quienes citan a Páramo y Correa (1999), se ha confirmado la hipótesis de que los estudiantes que abandonan sus estudios provienen principalmente de familias con recursos económicos limitados.

2.3.4. Factores académicos

Según la investigación de Tavico (2021), se destaca una estrecha relación entre los problemas académicos y el proceso educativo, siendo factores determinantes en la deserción escolar. El autor resalta la importancia de los procesos formativos previos de los estudiantes, especialmente en el ámbito de la educación media superior. Entre los aspectos significativos se incluyen el bajo rendimiento académico, la falta de preparación adecuada en el nivel educativo anterior (lo cual afecta negativamente el desempeño estudiantil), la escasa aplicación de técnicas de estudio apropiadas y los desafíos inherentes a la transición del nivel medio superior al nivel superior.

Los factores académicos también se relacionan con problemas cognitivos, como el bajo rendimiento académico, la repetición de materias, la falta de disciplina de los estudiantes y el uso de métodos de estudio obsoletos.

En este sentido se ha observado que los factores académicos son predominantes en la deserción escolar, ya que pueden llevar a los estudiantes a abandonar su formación debido a la pérdida de materias ocasionada por un bajo rendimiento académico (Ballesteros y Benalcázar, 2017).

2.3.5. Falta de apoyo de profesores e Insuficiente orientación académica

La desmotivación y la actitud negativa por parte de los estudiantes se han identificado como una causa importante de la deserción escolar. En este sentido, es fundamental que tanto las instituciones educativas con sus programas educativos como los docentes con sus prácticas pedagógicas desempeñen un papel crucial. Si bien es cierto que los problemas de actitud de los estudiantes pueden originarse en su desarrollo personal, también es cierto que es durante su trayectoria escolar donde se forja su desarrollo social. En otras palabras, todas las personas que forman parte de la comunidad educativa deben mostrar interés por los desafíos que enfrentan los estudiantes y abordarlos mediante diversas actividades que fomenten la integración y creen un ambiente propicio para el aprendizaje (Donoso, Donoso y Arias, 2018).

Además, se puede implementar programas de orientación vocacional y motivación específicos para la carrera que los estudiantes están cursando, lo cual tendría un impacto positivo tanto a nivel individual como en la sociedad en su conjunto. El propósito de estos programas es mejorar la actitud de las personas hacia el logro de sus metas, ya que esto es fundamental para alcanzar el éxito académico y personal (Ortega, 2012).

2.3.6. Problemas personales y familiares

El factor personal es otro aspecto fundamental para tener en cuenta, como menciona Zambrano (2021). Este factor abarca variables relacionadas con las características individuales de cada estudiante, sus metas, intereses personales, motivaciones y orientación vocacional. Estas variables influyen de manera significativa en la toma de decisiones académicas y pueden tener un impacto en la deserción escolar. Es importante

considerar y atender este aspecto para brindar un apoyo adecuado a los estudiantes y fomentar su compromiso y perseverancia en su proceso de formación.

Delors (1996) señala que, para abordar el tema de la deserción escolar, es fundamental identificar las desventajas que enfrentan los jóvenes, las cuales suelen estar relacionadas con su entorno familiar. Además, sugiere la implementación de políticas de discriminación positiva hacia aquellos estudiantes que presentan mayores dificultades.

Székely (2015), Weiss (2015) y Huerta, Velasco y Jiménez (2016) plantean que en el análisis de la deserción escolar es necesario tomar en cuenta la educación de los padres, así como su convivencia con sus hijos. También resaltan la importancia de considerar factores como la desmotivación, la falta de interés por la escuela, problemas relacionados con adicciones y la presencia de violencia en la vida de los estudiantes.

2.3.7. Desajuste con la vida universitaria

Algunos estudiantes pueden experimentar dificultades para adaptarse a la vida universitaria, especialmente si provienen de entornos diferentes o si tienen dificultades para relacionarse con otros estudiantes. El sentimiento de aislamiento social o la falta de sentido de pertenencia pueden contribuir a la deserción. Estudios recientes han concluido que el tema de la deserción universitaria se presenta principalmente en los primeros años de la carrera (Lugo, 2013).

Con relación a este tema, Tinto (2006) sostiene que hay dos períodos críticos en los que el riesgo de deserción es más elevado. El primero ocurre cuando el estudiante tiene su primer contacto con la institución y se forma una impresión inicial sobre sus características.

2.3.8. Problemas de salud mental

Según Sánchez *et al.,* (2009), en el ámbito personal, se involucran factores como los motivos psicológicos, que abarcan aspectos emocionales, motivacionales, desadaptación e insatisfacción. También se consideran los motivos sociológicos, que incluyen influencias familiares y de grupos como amigos o vecinos.

Además, Rochin Berumen (2021) destaca que, si se logra un cambio positivo en la mentalidad de los alumnos y se toma en cuenta su desarrollo individual, ellos mismos tendrán las herramientas para enfrentar no solo los desafíos académicos, sino también los personales.

2.3.9. Falta de vocación o elección de la carrera como última opción

Diversos autores (Canales y De los Ríos, 2007; Girón y González, 2005; Vries *et al.,* 2011) identificaron factores de deserción: desempeño escolar, apoyo familiar (Esteban *et al.,* 2017) constataron que el bajo rendimiento académico provoca el 45% del abandono en los dos primeros años. También el no ingresar a la carrera de primera opción (Abarca y Sánchez, 2005) y la procrastinación (Garzón y Gil, 2017) representan fuertes detonantes para el abandono.

2.3.10. Machine Learning

El concepto de "inteligencia artificial" se refiere a la capacidad de las máquinas para imitar funciones cognitivas que son características de la mente humana, por ejemplo, la percepción de datos, su razonamiento y método de implementación para la resolución de problemas. “La inteligencia artificial es un área de la investigación donde se desarrollan algoritmos para controlar cosas.” (Ino, 2008).

Dentro del campo de la inteligencia artificial se encuentra el aprendizaje automático (machine learning).

Machine learning, también conocido como aprendizaje automático en español, representa una aplicación de la inteligencia artificial (IA) que capacita a los sistemas para aprender de forma automática y mejorar con la experiencia, sin requerir una programación específica para ello. A lo largo del tiempo, el enfoque del aprendizaje automático ha evolucionado hacia el razonamiento probabilístico y la investigación basada en estadísticas. El proceso de aprendizaje inicia con la recopilación de datos u observaciones, y a partir de estas variables de entrada y sus respuestas se buscan patrones que permitan tomar decisiones futuras. Estas decisiones se basan en los datos

recopilados, lo que posibilita que las computadoras aprendan de manera automática sin necesidad de intervención humana (García Ruiz de León, 2018).

El término "aprendizaje automático" se refiere a la capacidad de detectar patrones de manera automatizada en los datos. En las últimas dos décadas, se ha convertido en una herramienta común en diversas tareas que involucran la extracción de información de conjuntos de datos extensos. Esta tecnología está presente en nuestra vida cotidiana: los motores de búsqueda aprenden para proporcionarnos los mejores resultados y mostrar anuncios relevantes, el software antispam aprende a filtrar nuestros correos electrónicos, y los sistemas de protección contra fraudes en transacciones con tarjeta de crédito emplean software que se entrena para detectar patrones sospechosos (Murphy, K., 2012).

Según Lugo (2020), el Machine Learning se define como la disciplina científica que posibilita que las computadoras aprendan y emulen las acciones humanas, mejorando su capacidad de aprendizaje de manera autónoma a lo largo del tiempo mediante la alimentación de datos e información en forma de observaciones e interacciones con el mundo real. En este sentido, Fagella (2020) expone que la tecnología de inteligencia artificial se emplea en diversas aplicaciones, destacando su utilización en el sector minorista, donde se implementa como una innovación a lo largo de todo el ciclo de productos y servicios.

Por otra parte, el ML se basa en algoritmos de Inteligencia Artificial (IA). De acuerdo con Arbeláez-Campillo, Villasmil y Rojas-Bahamón (2021), la IA, respaldada por información cualitativa y cuantitativa eficiente, tiene la capacidad de alcanzar diversos objetivos. En este caso, los algoritmos de Machine Learning tienen como objetivo identificar patrones a partir de los datos recopilados y, posteriormente, utilizar un modelo para realizar predicciones. Para lograr esto, se pueden emplear datos tanto estructurados como no estructurados. (Digital55 2020).

2.3.11. Tipos de Aprendizaje Automático

Los modelos de aprendizaje automático se fundamentan principalmente en tres tipos de aprendizaje:

Aprendizaje supervisado. Gironés et al. (2017) y Kubat (2017) explican que el aprendizaje supervisado abarca diversos algoritmos que, durante su fase de aprendizaje, dependen de un conjunto de datos de entrenamiento donde se conocen los valores o clases de una variable objetivo o de respuesta. En este proceso, el algoritmo realiza predicciones para la variable objetivo basándose en un conjunto de variables predictoras X cuyos valores son conocidos. La denominación de aprendizaje supervisado se debe a que las variables X supervisan la respuesta Y.

Aprendizaje no supervisado. Este enfoque se emplea cuando los datos no tienen clasificación y carecen de cualquier indicación preexistente. En esta situación, los datos de entrada se consideran variables aleatorias, y el sistema debe identificar patrones para clasificar nuevas entradas sin la guía de etiquetas previas (García Ruiz de León, 2018).

Aprendizaje por refuerzo o semi-supervisado. El aprendizaje semi-supervisado, según Xiaojin (2008), hace uso de una extensa cantidad de datos, algunos de los cuales pueden tener o no etiquetas conocidas. Esta metodología tiene como objetivo reducir las limitaciones y mejorar la construcción de clasificadores. Por ende, el aprendizaje semi-supervisado implica una intervención humana menor y está diseñado para ser más flexible, proporcionando una mayor precisión en sus predicciones.

En este estudio, se dará prioridad al aprendizaje supervisado para desarrollar nuestros modelos, este enfoque es extensamente utilizado en contextos estadísticos que involucran regresión y predicción, enfocándose en el uso de datos previamente etiquetados para orientar al sistema en su proceso de aprendizaje.

Según Gironés *et al.*, (2017), el aprendizaje supervisado se divide en dos tareas esenciales: regresión y clasificación. En la regresión, que pertenece al ámbito del aprendizaje automático, la variable objetivo a predecir es de naturaleza numérica. Por

otro lado, la clasificación supervisada se emplea cuando la variable objetivo a predecir es de carácter categórico.

2.3.12. Tipos de algoritmos de clasificación de aprendizaje automático

Con el objetivo de abordar la problemática de clasificación supervisada para la creación de un modelo predictivo de deserción escolar, se evaluarán y aplicarán cuidadosamente diferentes algoritmos de Machine Learning. Estas elecciones se han considerado meticulosamente con el propósito de ofrecer soluciones precisas a la cuestión planteada. Entre los muchos modelos de aprendizaje automático, vamos a destacar los siguientes:

Algoritmo Clasificador bayesiano (NaiveBayes)

Un clasificador Bayesiano ingenuo es un clasificador probabilístico simple basado en el teorema de Bayes con fuertes supuestos (ingenuos) de independencia de características. En términos sencillos, un clasificador Bayesiano ingenuo asume que la presencia o ausencia de una característica concreta no está relacionada con la presencia o ausencia de cualquier otra característica, dada la variable de clase, (John y Langley, 1995).

De acuerdo con (Báez, 2016), este algoritmo es altamente relevantes ya que permiten un análisis cualitativo y cuantitativo de los atributos y valores que influyen

en un problema, esto las convierte en un método significativo en tareas como la clasificación de documentos o la creación de filtros de correo electrónico, al igual que los árboles de decisión y las redes neuronales artificiales, que han sido los tres métodos más utilizados en aprendizaje automático en los últimos años.

El algoritmo Naive Bayes, ha sido utilizado por muchos estudios desarrollando modelos sobre este tema de la deserción, Agrawal *et al.,* (2015) afirma que el algoritmo Naive Bayes tiene un buen rendimiento en la predicción del rendimiento de los estudiantes.

Arboles de Decisión para Clasificación

Se encuentra clasificado dentro de los modelos de predicción, se fundamenta en el aprendizaje inductivo partiendo de observaciones lógicas, el cual tiene un comportamiento en sus predicciones basadas en reglas que permiten la representación y categorización de las condiciones que se suscitan de forma repetitiva para la solución

del problema. Su representación gráfica se hace por medio de un conjunto de nodos, hojas y ramas (Barrientos M, 2009). Este método de clasificación no almacena los datos de entrenamiento en su totalidad, utiliza los datos de entrenamiento para construir una estructura de estratificación o de árbol que divida recursivamente el espacio de entrenamiento en regiones que tengan etiquetas o características similares. Tiene la característica de realizar una partición recursiva en donde la variable dependiente puede ser continua, discreta o categórica y las variables predictoras pueden ser continuas, discretas o categóricas y finalmente, tiene la facilidad de poseer el esquema de interpretabilidad. (Ramasubramanian, 2019).

Máquinas de Soporte Vectorial (SVM)

El algoritmo Máquinas de Soporte Vectorial, del inglés *Support Vector Machine (SVM),* tiene como objetivo encontrar el hiperplano más favorable que distinga eficazmente las muestras pertenecientes a diferentes clases dentro de un espacio multidimensional. SVM se esfuerza por identificar el hiperplano que maximiza el margen entre estas clases, mejorando así su capacidad para manejar nuevos puntos de datos (Probst *et al.,* 2018). Cuando se enfrenta a clases no linealmente separables, utiliza funciones de núcleo para transformar los datos en un espacio de mayor dimensión en el que se pueden distinguir. De este modo, el algoritmo puede clasificar nuevas instancias en función de su posición relativa al hiperplano y asignarlas con precisión a las clases predefinidas adecuadas.

XGBoost

XGBoost, Chen *et al.* (2016) es un sistema de boosting de árboles propuesto por Friedman (2001) comúnmente utilizado por los científicos de datos para lograr mejores resultados en problemas de aprendizaje automático. Se caracteriza por ser rápido, preciso, flexible y eficiente debido a que está diseñado según Chen *et al.* (2016) para consumir una menor cantidad de recursos por lo que obtiene buenos resultados con mínimo esfuerzo.

Algoritmo KNN (K-Nearest Neighbor)

K-Nearest Neighbor podría describirse como aprendizaje por analogía, ya que se aprende comparando una tupla de prueba específica con un conjunto de tuplas de entrenamiento

que son similares a ella. Se clasifica en función de la clase de sus vecinos más cercanos. Por lo general, se tiene en cuenta a más de un vecino, de ahí el nombre de K-Nearest Neighbor (K-NN). La «K» indica el número de vecinos que se tienen en cuenta para determinar la clase (Han & Kamber, 2006). El algoritmo K-NN ha sido adoptado por los estadísticos como un enfoque de aprendizaje automático hace más de 50 años, (Markov & Larose, 2007). El algoritmo K-NN suele denominarse «aprendiz perezoso» en el sentido de que simplemente almacena las tuplas de entrenamiento dadas y espera hasta que se le proporciona una tupla de prueba, momento en el que realiza la generalización para clasificarla basándose en similitudes o distancias con las tuplas de entrenamiento almacenadas. También se denomina «aprendiz basado en instancias». El aprendiz perezoso o basado en instancias realiza menos trabajo cuando se le presentan tuplas de entrenamiento y más trabajo durante la clasificación y la predicción, por lo que es caro computacionalmente, a diferencia de los aprendices ansiosos que, cuando se les da una tupla de entrenamiento, construyen un modelo de clasificación antes de recibir la tupla de prueba para clasificar, por lo que están muy preparados y ansiosos por clasificar cualquier tupla no vista. Peterson (2008), junto con Yoti y Walia (2017), Amartya y Kundan (2007), así como Han y Kamber (2006) y Markov y Larose (2007), afirman que el error K-NN está acotado por encima del doble de la tasa de error de Bayes.

2.3.13. Técnicas de selección de características

El uso de algoritmos de aprendizaje automático en conjuntos de datos con un gran número de atributos puede generar varios problemas que afectan significativamente a su rendimiento. Estos problemas incluyen el sobreajuste, el aumento de los tiempos de cálculo y aprendizaje, así como la degradación del modelo en presencia de datos ruidosos.

Para hacer frente a los problemas mencionados, una de las herramientas más eficaces es la reducción de la dimensionalidad de los conjuntos de datos. Esta técnica se basa en la selección de un subconjunto de características que contenga la información más

relevante (Alonso-betanzos, 2007). Al trabajar con un conjunto de datos optimizado con características significativas, es posible, según Li *et al* (2018), mejorar significativamente el rendimiento predictivo de un modelo de aprendizaje automático, reducir la complejidad del modelo, disminuir los costes computacionales y de recursos, y evitar el sobreajuste de los algoritmos.

Aunque los expertos en la materia pueden identificar y eliminar algunos atributos irrelevantes, la selección óptima de un subconjunto de características suele requerir un planteamiento más estructurado y sistemático. Para hacer frente a este reto, existen tres grandes familias de métodos automáticos de selección de características:

Métodos de filtrado. suelen utilizarse como paso previo al procesamiento. La selección de características es independiente de cualquier algoritmo de aprendizaje automático y se basa en puntuaciones obtenidas a partir de diversas pruebas estadísticas (Talavera, 2005). Ejemplos.

- Chi-cuadrado: Evalúa la independencia entre las características y la variable objetivo.
- Información mutua: Mide la cantidad de información compartida entre las características y la variable objetivo.
- Puntuación F: Utiliza el estadístico F para evaluar la importancia de las características en los problemas de clasificación.
- Valor informativo: Evalúa la capacidad predictiva de una característica en relación con la variable objetivo.

Métodos de envoltura (wrappers). Estos métodos utilizan un subconjunto de características para construir un modelo y, en función del rendimiento del modelo, se decide si añadir o eliminar características de este subconjunto (Karegowda *et al.,* 2010). Ejemplos.

- Selección secuencial hacia delante (SFS): añade características una a una, seleccionando la que más mejora el modelo en cada paso.

- Selección secuencial hacia atrás (SBS): comienza con todas las características y elimina la menos significativa en cada paso.
- Búsqueda recursiva de características (RFE): Selecciona características eliminando iterativamente las que menos afectan al rendimiento del modelo.

Algoritmos genéticos. Utilizan técnicas evolutivas para explorar combinaciones de características que optimicen el rendimiento del modelo.

- Métodos embebidos (embedded): estos métodos seleccionan características durante la ejecución del algoritmo de aprendizaje, integrándose como parte del proceso de aprendizaje, (Jović *et al.*, 2015). Ejemplos.
- LASSO (Least Absolute Shrinkage and Selection Operator): realiza una regularización que obliga a que algunos coeficientes de características sean exactamente cero, eliminándolos del modelo.
- Árboles de decisión: Seleccionan características en función de su capacidad para dividir los datos en nodos homogéneos.
- Máquinas de vectores soporte (SVM) con selección de características: Incorporan la selección de características en el proceso de optimización de márgenes.
- Redes neuronales con regularización: Utilizar técnicas de regularización, como la penalización L1, para reducir el número de características relevantes.

2.3.14. Métricas de evaluación en modelos de clasificación

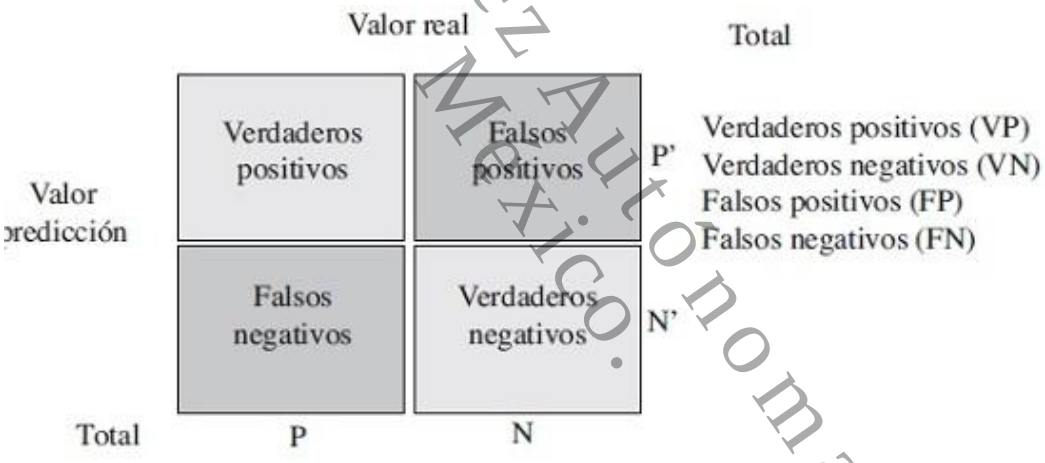
La matriz de confusión, también conocida como matriz de clasificación, incluye cuatro métricas clave para evaluar el desempeño de un modelo: Falsos Positivos (FP), Falsos Negativos (FN), Verdaderos Positivos (VP) y Verdaderos Negativos (VN), (S. Kotsiantis *et al.,* 2004). Entre las métricas de rendimiento esta la exactitud, precisión, métrica de exhaustividad (Recall), F1-Score. Todas estas métricas se obtienen de Scikit-learn la cual es una herramienta con algoritmos de aprendizaje de Machine Learning que ayudan en el momento de evaluar modelos.

Para entender cómo funcionan las métricas de evaluación se utiliza la matriz de confusión. Una matriz de confusión se representa como una matriz de 2 x 2 que da las predicciones de verdadero positivo, verdadero negativo, falso positivo y falso negativo de un clasificador de imágenes, esta matriz nos da información de los posibles errores que se pueden cometer en la clasificación, por lo tanto, una matriz de confusión ayuda a evaluar la exactitud o precisión del clasificador. (Zapeta *et al.,* 2022).

En la figura 1 se muestra los cuadrantes que se utilizan en la matriz de confusión para realizar la predicción.

Figura 1

Matrix de confusión



Nota: Fuente: Universidad Externado de Colombia.

a. **Verdadero positivo**: Se refiere al caso en que el valor real es positivo y el modelo acierta al predecir que también es positivo.
b. **Verdadero negativo**: Ocurre cuando el valor real es negativo y el modelo correctamente identifica que es negativo.

c. **Falso negativo**: Se da cuando el valor real es positivo, pero el modelo erróneamente predice que es negativo.
d. **Falso positivo**: Sucede cuando el valor real es negativo, pero el modelo incorrectamente clasifica el valor como positivo.

Precisión

El rendimiento de los modelos de clasificación se evalúa cuantitativamente usando métricas derivadas de la matriz de confusión, como Verdaderos Positivos, Verdaderos Negativos, Falsos Positivos y Falsos Negativos.

Estas métricas básicas permiten calcular otras métricas adicionales que proporcionan una visión más detallada del desempeño del modelo, como Precisión de clasificación es lo que normalmente se dice cuando se usa el término precisión. Es la conexión entre el número de predicciones correctas y el número total de muestras de entrada (Mishra, 2018).

$$Presicion = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FP}}$$

Recall

La métrica sensibilidad, mejor conocida como Sensitivity/Recall, representa la relación entre los casos positivos identificados correctamente y el número total de casos positivos reales. Mide la capacidad del modelo para encontrar todos los casos positivos. Razón por la que también es denominada fracción de verdaderos positivos (FVP), (López de Ullibarri & Píta Fernández, 1998).

$$Recall = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FN}}$$

Specificity

La métrica especificad, mejor conocida como *Specificity*, mide la proporción de verdaderos negativos identificados correctamente sobre el número total de negativos reales. Evalúa la capacidad del modelo para identificar correctamente los casos sin riesgo

$$Recall = \frac{\text{VP}}{\text{VN} + \text{FP}}$$

Puntaje F1

La métrica puntaje F1, mejor conocida como F1-Score, combina la métrica Precisión y Sensibilidad en una única métrica que representa el equilibrio entre ambas. Resulta útil cuando se desea tener en cuenta tanto la precisión como la sensibilidad en la evaluación de modelos.

$$F - Score = \frac{2 \text{ x Precision x Sensibilidad}}{\text{Precision} + \text{Sensibilidad}}$$

2.3.15. Minería de Datos

Según Aquino, Molero y Rojano (2015), "La minería de datos es un dominio de la ciencia de la computación que permite el análisis de grandes cantidades de datos para encontrar y extraer patrones significativos útiles para el proceso de toma de decisiones". Además, advierten que la minería de datos va más allá de simples tareas como la revisión de los datos en bases de datos, está más encaminado hacia el análisis de grandes volúmenes de información para encontrar patrones relevantes en diferentes áreas de la ciencia como educación, medicina, finanzas etc.

Tipos de Análisis

En minería de datos se encuentran básicamente 2 tipos de análisis: El descriptivo y el predictivo, estos permiten desplegar diferentes tipos de tareas como la clasificación (Kumari *et al., 2011*), la predicción (Shaikh T, 2014) la segmentación (Oviedo *et al., 2015*) y la asociación (Amin A, T. *et al., 2014*).

En el análisis descriptivo las aplicaciones más comunes son: Detección de anomalías, Análisis de perfil de personas, detección de reglas que condicionen la venta de productos, etc. (*Riquelme J et al., 2006*). En este análisis, el conjunto de datos está conformado por los atributos que se desean analizar para encontrar patrones o asociaciones de los datos. Se pueden desarrollar tareas de agrupación (*clustering*) y de asociación (Oviedo *et al.,* 2015).

El análisis predictivo se caracteriza por el uso de un conjunto de entrenamiento que está formado por datos históricos. Las aplicaciones comunes son predecir riesgos, predicción en las ventas, activación de nuevos clientes, etc. (Riquelme J *et al*., 2006). En este análisis se desarrollan tareas de predicción continua (numérica) y predicción discreta (clasificación) (Oviedo *et al.,* 2015).

Técnicas de minería de datos

En general las técnicas se pueden agrupar en técnicas supervisadas y no supervisadas (Oviedo et al., 2015). Las técnicas supervisadas son usadas para análisis predictivo, algunas de estas técnicas son: Arboles de decisión, Redes Neuronales, Métodos de regresión, Máquinas de soporte vectorial, Métodos basados en ejemplos y Método Bayesiano (Oviedo et al., 2015).

Las técnicas no supervisadas son usadas en el análisis descriptivo, algunas técnicas son: Método Jerárquico, Método Particional, Redes Neuronales, Reglas de Asociación y Método Probabilístico (Oviedo et al., 2015).

2.4. Marco Tecnológico

En la actualidad, existen diversas herramientas tecnológicas que facilitan el desarrollo de modelos predictivos de aprendizaje automático. Estas herramientas proporcionan diferentes funcionalidades que abarcan desde la manipulación y visualización de datos

hasta el entrenamiento y la evaluación de modelos. A continuación, se describen algunas de las herramientas más destacadas, junto con sus características:

Lenguaje de Programación y librerías

Como principal herramienta de desarrollo se usará el lenguaje de programación Python un lenguaje flexible de alto nivel que es capaz de ser usado en diversas tareas desde programación web hasta análisis de datos, construcción de algoritmos de redes neuronales y de minería de datos, su curva de aprendizaje es menor a la de otros usado en trabajos relacionados como el lenguaje R, M y Java, lo que facilita la interpretación de algoritmos en código.

Python

Python fue creado a principio de los años 90 por Guido van Rossum en los Países Bajos como sucesor del lenguaje ABC. En el 2001 fue creada *Python Software Foundation*, organización sin ánimo de lucro que actualmente es la que se encarga oficialmente de la coordinación del desarrollo, mantenimiento y documentación del lenguaje. (*Python Software Foundation* 2021).

NumPy

NumPy es una librería de Python que te permite trabajar con grandes conjuntos de datos numéricos de manera eficiente, con NumPy, puedes crear y manipular arrays y matrices, lo que significa que puedes realizar cálculos complejos en muy poco tiempo.

Pandas

Pandas es una biblioteca de Python que ofrece estructuras de datos eficientes, adaptables y expresivas diseñadas para facilitar el manejo de datos "relacionales" o "etiquetados". Su propósito fundamental es servir como un componente clave de alto nivel

para llevar a cabo análisis de datos prácticos y del mundo real de manera fácil e intuitiva en el entorno de programación Python.

Scikit-learn

Scikit-learn es una biblioteca de código abierto para Python que facilita el desarrollo de modelos de aprendizaje automático. Construida sobre NumPy, la biblioteca proporciona una amplia gama de algoritmos avanzados, incluyendo KNN, XGBoost, bosques aleatorios y máquinas de vectores de soporte (SVM), entre otros, tambien ofrece herramientas para el preprocesamiento de datos, la reducción de dimensionalidad (como la selección de características), la clasificación, la regresión, y la agrupación. Además, es compatible con otras librerías de Python como SciPy y matplotlib, lo que la convierte en una herramienta integral para el análisis y modelado de datos.

2.4.1. Entorno de Desarrollo Integrado (IDE)

Google Colab

Colab es un servicio alojado de Jupyter Notebook que no requiere configuración y que ofrece acceso sin coste económico a recursos de computación, como GPUs y TPUs. Es una solución especialmente adecuada para el aprendizaje automático, la ciencia de datos y la educación.

2.5. Marco legal

2.5.1. Licencia de Software Libre

Para el desarrollo del presente trabajo se utilizaron versiones de software y plataformas de código abierto, así como herramientas de desarrollo de acceso público. Entre estas, se incluyen el entorno de desarrollo integrado (IDE) Google Colab, lenguajes de programación como Python y librerías como scikit-learn, numpy y pandas, las cuales no requieren licencias comerciales para su utilización. Se garantizará el cumplimiento de las

licencias de código abierto seleccionadas, respetando los términos y condiciones establecidos por las respectivas comunidades de desarrollo y los acuerdos de uso vigentes.

Capítulo III. Aplicación de la Metodología

3.1. Muestreo

La definición de muestreo es muy práctica y comprensible gracias al manejo conceptual que se ha venido acuñando en las diversas disciplinas de la ciencia. Por eso es necesario especificar con más certeza el significado del término para poder comprender su aplicación en el ámbito de la investigación. Por este motivo, se expondrán dos posturas teóricas: la de Ander-Egg (2006) y la de Buendía, Colas y Hernández (1999), que especifican con más precisión las características que debe tener el término. El primero establece que la población, o, en términos más precisos, población objetivo, es un conjunto finito o infinito de elementos con características comunes para los cuales se extenderán las conclusiones de la investigación. Esta queda delimitada por el problema y por los objetivos del estudio (Ander-Egg, 2006). El segundo parte de elementos, personas o fenómenos que constituyen la muestra de la investigación, y viene acompañada por un grupo de conceptos básicos que conviene clarificar: universo, población, muestra, individuo, etc. (Buendía, Colás y Hernández, 1999).

Muestreo estratificado

El muestreo estratificado es un diseño de muestreo probabilístico en el que dividimos la población en subgrupos o estratos. La estratificación puede basarse en una amplia variedad de atributos o características de la población como edad, género, nivel socioeconómico, ocupación, etc.

Para este estudio, se empleará un muestreo estratificado con un enfoque cuantitativo, ya que el objetivo es obtener un conjunto de datos estadísticamente significativo que permita identificar patrones generales de deserción a través de algoritmos de aprendizaje automático. La técnica específica de muestreo será probabilística, en concreto, muestreo aleatorio estratificado. Esta técnica se elige para asegurar que la muestra represente de manera adecuada las características distintivas de cada carrera o área de estudio,

minimizando posibles sesgos y permitiendo una evaluación precisa de los factores que influyen en la deserción.

3.2. Población estudiada

La población de este estudio se compone de estudiantes universitarios de diversas áreas de conocimiento que están matriculados en los primeros semestres de sus carreras. Este grupo es relevante para la investigación, ya que la deserción tiende a ser más común en las primeras etapas del ciclo universitario, donde se presentan mayores retos de adaptación y dificultades académicas. Además, incluir estudiantes de distintos campos académicos permitirá obtener un análisis más completo y generalizable.

3.3. Diseño Experimental

Diseño no experimental

Dado que no se manipularán activamente las variables (por ejemplo, no se intervendrá sobre los estudiantes para alterar factores que influyan en la deserción), el estudio será observacional. El análisis se basará en datos existentes sobre estudiantes (demográficos, académicos, socioeconómicos, etc.), y no en la aplicación de tratamientos o intervenciones experimentales.

3.4. Procedimiento

Metodología de minería de datos en el desarrollo de modelos de aprendizaje automático

Metodología KDD

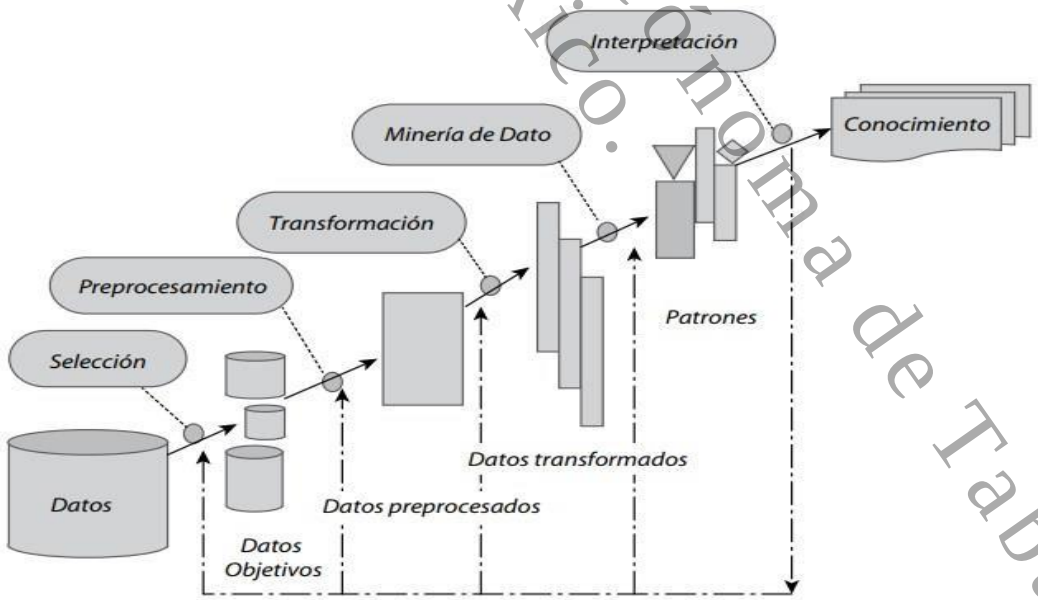
Knowledge Discovery in Database (KDD) es básicamente un proceso automático en el que se combinan el descubrimiento y el análisis de datos. Este proceso implica extraer patrones en forma de reglas o funciones de los datos para el análisis del usuario. (Timarán-Pereira *et al.,* 2016).

En el desarrollo de esta propuesta tecnológica se empleó el método de Descubrimiento de Conocimiento en Bases de Datos (KDD), debido a las características iterativas e interactivas de sus distintas etapas, tal como se observa en la Figura 2.

Este enfoque es iterativo, ya que permite volver a pasos previos cuando los resultados de una fase específica requieren ajustes o mejoras en las etapas anteriores. Este proceso cíclico es esencial para alcanzar un conocimiento de alta calidad, ya que, en la práctica, suelen ser necesarias múltiples iteraciones para extraer patrones significativos y relevantes de los datos. Además, el método KDD es interactivo, lo que implica la participación de los usuarios en la toma de decisiones a lo largo del proceso, de modo que sus conocimientos y criterios orienten las fases críticas de la extracción de información y el análisis de datos. Esta interacción garantiza que el conocimiento generado sea relevante y esté alineado con los objetivos específicos del estudio.

Figura 2

Metodología KDD



Nota: Tomado de Pereira, Arteaga, Caicedo, Hidalgo y Pérez 2016

Etapas de la Metodología KDD

Selección

En esta etapa, se crea o se busca el conjunto de datos de destino, se selecciona todo el conjunto de datos o una muestra representativa del conjunto de datos y luego se realiza el proceso de descubrimiento en el conjunto de datos. La elección de datos varía según los objetivos del negocio. (Timarán-Pereira *et al.,* 2016). El primer paso para extraer conocimiento de los datos es identificar y recopilar los datos que se utilizarán. Para ello se identificará la base de datos integrada al sistema, así como los datos y usos que conforman la base de datos.

Preprocesamiento/limpieza

En la fase de preprocesamiento/limpieza, se analizan la calidad de los datos, se aplicarán operaciones básicas, como eliminar datos ruidosos, elegir estrategias para lidiar con datos desconocidos, datos vacíos, datos duplicados y técnicas estadísticas alternativas para su reemplazo. (Timarán-Pereira *et al.,* 2016). En esta etapa se analizarán las tablas que integran la base de datos del sistema y se seleccionarán aquellas que se consideren necesarias para la optimización del algoritmo.

Transformación/reducción

En la etapa de transformación/reducción de datos, se buscan características útiles para representar los datos de acuerdo con los objetivos del proceso. (Timarán-Pereira *et al.,* 2016). Los métodos de transformación o reducción de dimensionalidad se utilizan para reducir el número efectivo de variables consideradas o para encontrar representaciones invariantes de datos. (Timarán-Pereira et al, 2016). En esta etapa se realiza la selección de los atributos requeridos, para utilizarlos en los algoritmos.

Minería de datos (data mining)

En la fase de minería de datos se aplicarán modelos, tareas, técnicas y algoritmos seleccionados para obtener reglas y patrones. (Brito Sarasa *et al.,* 2008). En esta etapa, se seleccionan y aplican técnicas de minería de datos adecuadas, que puedan cumplir con los objetivos propuestos, para lo cual se recopila teóricamente la información necesaria.

Utilizando la tecnología adecuada, se procederá con el análisis manipulando losl algoritmo de aprendizaje automático para obtener patrones que permita dar una clasificación de líneas y sub líneas de investigación en base a los atributos previamente seleccionados. Es importe recalcar que para realizar este proceso se utiliza el lenguaje de programación Python.

Interpretación/evaluación

En esta etapa se explicarán los patrones descubiertos y se podrá devolver la etapa anterior para iteraciones posteriores, pudiendo también incluir la visualización de los patrones extraídos. Por otro lado, se consolida el conocimiento descubierto para fusionarlo en otro sistema para operaciones posteriores, o simplemente registrarlo e informar a las partes relevantes; también puede verificar y resolver conflictos potenciales del conocimiento previamente descubierto. (Timarán-Pereira *et al.,* 2016).

En esta etapa se evaluarán los resultados obtenidos mediante la aplicación del algoritmo y se verificará si cumple con los objetivos de predicción y clasificación de líneas y sub líneas de investigación, y luego se incluirá en el sistema.

Validación cruzada

En el ámbito de Inteligencia Artificial existen varios métodos que pueden ser utilizados para realizar el proceso de validaciones. Sin embargo, muchos expertos consideran que el método de Validación Cruzada o *Cross-Validation* es el más recomendado para realizar

el análisis del nivel de exactitud que un algoritmo posee. Según Pérez-Planells., *et al* (2016), la validación cruzada es una técnica que se utiliza para evaluar los resultados del análisis estadístico y garantizar que sea independiente de la división entre los datos de entrenamiento y los datos de prueba.

3.5. Análisis de los Datos

El análisis de datos es el proceso de examinar, limpiar, transformar y modelar datos para descubrir información útil, sacar conclusiones y respaldar la toma de decisiones. (Tukey, 1962).

Como menciona Rodríguez (2023), el aprendizaje automático y el análisis de datos emplean diversas metodologías y modelos para identificar patrones ocultos y relaciones subyacentes en los conjuntos de datos, permitiendo realizar predicciones o tomar decisiones informadas. En esta sección, se describirán algunas de las técnicas y algoritmos y pasos que se llevarán a cabo en esta investigación.

Recolección y comprensión de los datos

El primer paso consistirá en familiarizarse con el conjunto de datos disponible, que incluye información demográfica, académica y socioeconómica de los estudiantes. Se revisarán atributos como la edad, género, situación económica, y desempeño académico, se verificará la calidad de los datos, identificando y manejando valores faltantes, duplicados o inconsistentes. También se normalizarán los valores categóricos si es necesario.

Exploración inicial de los datos (EDA)

Se realizaron análisis descriptivos, como histogramas, gráficos de barras y diagramas de dispersión, para entender la distribución de cada variable. Esto incluirá la observación de la edad de los estudiantes, el número de unidades curriculares completadas, entre otras variables, se explorarán correlaciones entre las características del *dataset* para identificar

relaciones significativas, como la influencia de la situación socioeconómica en el rendimiento académico.

Selección de características

Se emplearon las técnicas de selección de características como *ANOVA, Mutual Información, Regresión Lasso, y Recursive Feature Elimination (RFE)* para reducir la cantidad de características, seleccionando solo aquellas que aporten mayor relevancia al modelo predictivo.

Estas técnicas permitirán identificar las variables más influyentes en la predicción de deserción, como el estado de matrícula, becas recibidas, ocupación de los padres, y rendimiento académico en los primeros semestres.

Transformación y balanceo de los datos

Para abordar posibles desbalances en las clases (por ejemplo, mayor cantidad de estudiantes que no desertan), se aplicará la técnica de balanceo SMOTEENN, que genera nuevos ejemplos sintéticos y elimina ruido, garantizando una distribución equitativa en las clases objetivo, después el conjunto de datos será dividido en entrenamiento y prueba, asegurando que la evaluación del modelo se realice de manera imparcial.

Entrenamiento y evaluación de modelos

Se entrenarán diversos modelos de aprendizaje automático, incluyendo *K-Nearest Neighbors* (KNN), Árboles de Decisión (DT) y Máquinas de Vectores de Soporte (SVM), posteriormente se medirán métricas como precisión, exactitud, *recall,* y *F1-score* para cada modelo y técnica de selección de características. Esto permitirá identificar el mejor modelo y la técnica más adecuada para predecir la deserción.

Optimización y validación del modelo

Los modelos se afinarán mediante la búsqueda de los mejores hiperparámetros utilizando técnicas como Grid Search o Random Search, se realizará una validación cruzada para garantizar que el modelo no esté sobre ajustado y que tenga un buen desempeño en datos no vistos.

Interpretación de resultados

Finalmente, se interpretarán los resultados obtenidos en términos de las variables más influyentes y se analizarán las implicaciones de los hallazgos para las instituciones educativas. Se destacará cómo el modelo puede ser utilizado para identificar de manera temprana a los estudiantes en riesgo de abandonar sus estudios y qué factores tienen mayor impacto en la deserción.

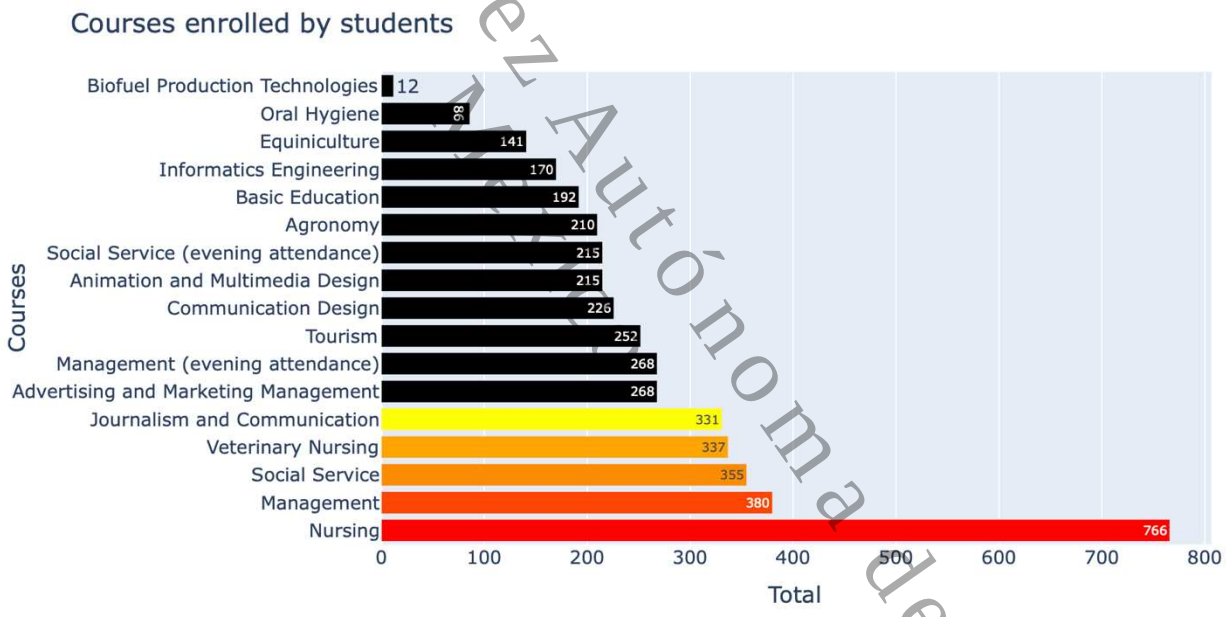
Capítulo IV. Resultados y Discusión

4.1. Recopilación de datos

En este estudio se utilizó el conjunto de datos *Predict Students' Dropout and Academic Success*, que contiene datos demográficos, macroeconómicos y socioeconómicos, junto con información sobre la matrícula de los estudiantes de los primeros y segundos semestres de 17 carreras universitarias de diversa índole. Véase la Figura 3.

Figura 3

Matricula por carrera universitaria



Nota: Elaboración propia con base en estudios de Realinho et al., (2021)

El conjunto de datos analizado consta de 4424 instancias y 34 atributos, además de una variable objetivo multiclase, de estos, 20 son discretos, 8 binarios, 5 continuos y 1 ordinal. Cada tipo de atributo cumple una función específica en la descripción de las características estudiantiles universitaria. Para más detalles sobre la estructura y los tipos de atributos incluidos, véase la tabla 4.

Tabla 4

Distribución de atributos y característica del conjunto de datos.

Class of Attribute	Attribute	Attribute Type
Demographic data	Marital status	Numeric/discrete
	Nationality	Numeric/discrete
	Displaced	Numeric/binary
	Gender	Numeric/binary
	Age at enrollment	Numeric/discrete
	International	Numeric/binary
Socioeconomic data	Mother's qualification	Numeric/discrete
	Father's qualification	Numeric/discrete
	Mother's occupation	Numeric/discrete
	Father's occupation	Numeric/discrete
	Educational special needs	Numeric/binary
	Debtor	Numeric/binary
	Tuition fees up to date	Numeric/binary
	Scholarship holder	Numeric/binary
Macroeconomic data	Unemployment rate	Numeric/continuous
	Inflation rate	Numeric/continuous
	GDP	Numeric/continuous
Academic data at enrollment	Application mode	Numeric/discrete
	Application order	Numeric/ordinal
	Course	Numeric/discrete
	Daytime/evening attendance	Numeric/binary
	Previous qualification	Numeric/discrete
Academic data at the end of 1st semester	Curricular units 1st sem (credited)	Numeric/discrete
	Curricular units 1st sem (enrolled)	Numeric/discrete
	Curricular units 1st sem (evaluations)	Numeric/discrete
	Curricular units 1st sem (approved)	Numeric/discrete
	Curricular units 1st sem (grade)	Numeric/continuous
	Curricular units 1st sem (without evaluations)	Numeric/discrete
Academic data at the end of 2nd semester	Curricular units 2nd sem (credited)	Numeric/discrete
	Curricular units 2nd sem (enrolled)	Numeric/discrete
	Curricular units 2nd sem (evaluations)	Numeric/discrete
	Curricular units 2nd sem (approved)	Numeric/discrete
	Curricular units 2nd sem (grade)	Numeric/continuous
	Curricular units 2nd sem (without evaluations)	Numeric/discrete

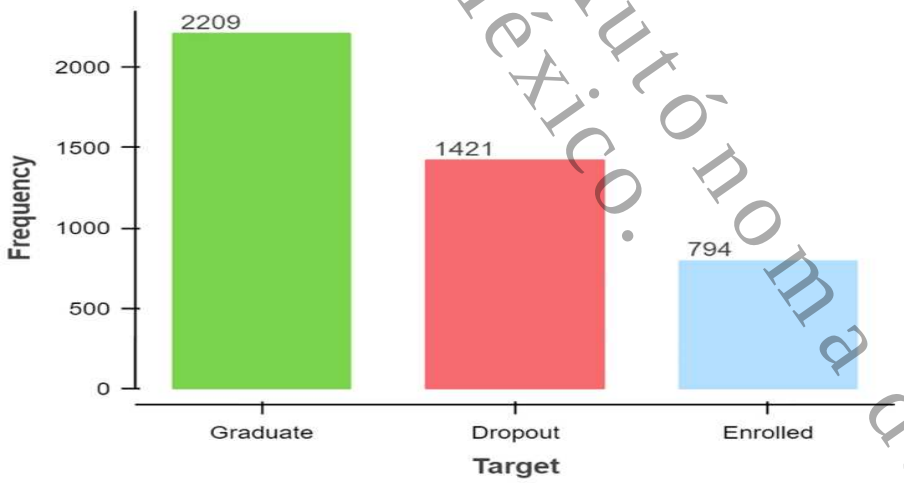
Nota: Elaboración propia con base en estudios de Realinho et al., (2021)

Análisis y visualización de los Datos

La variable objetivo tiene tres clases etiquetadas como Desertor (0), Matriculado (1) y Graduado (2). Como se observa en la figura 4, los datos están desequilibrados en lo que respecta a los alumnos matriculados: hay 1421 casos (32,1 %) de la clase etiquetada como «abandono», mientras que de la clase etiquetada como «matriculado» hay 794 casos (17,9 %). Por último, la clase etiquetada como «graduado» tiene 2209 instancias (49,9 %).

Figura 4

Distribución de las variables objetivos del conjunto de datos.



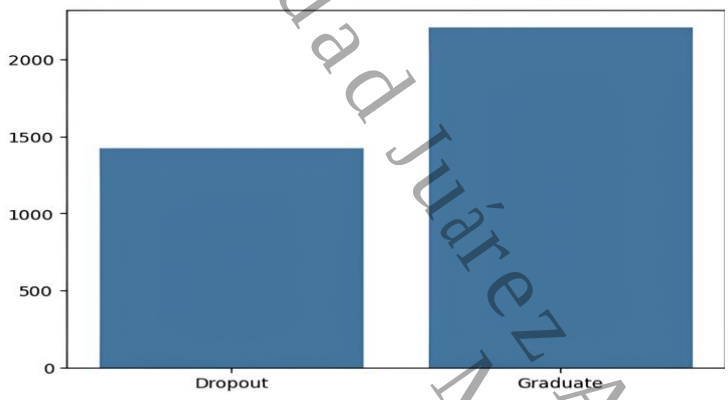
Nota: Elaboración propia con base en estudios de Realinho et al., (2021).

Se identificó que la variable objetivo «Matriculado» no aportaba un valor significativo para el análisis. Por esta razón, se optó por eliminarlos, de modo que el estudio se centrara

exclusivamente en las clases Desertor y Graduado, como resultado, el conjunto de datos final se redujo a 3,630 registros. Ver figura 5.

Figura 5

Distribución de las variables objetivos después de eliminar la variable matriculada.

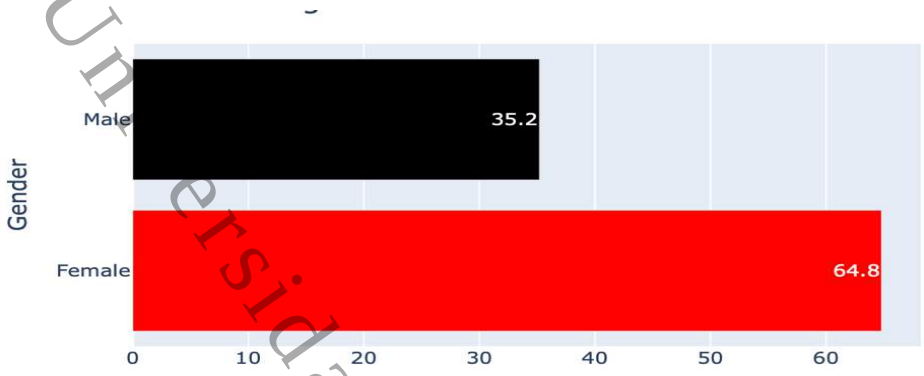


Nota: Elaboración propia con base en estudios de Realinho et al., (2021).

El análisis reveló un desequilibrio de género en el conjunto de datos, con un 64.8% de estudiantes mujeres y un 35.2% de hombres. Además, la tasa de deserción es más alta entre los hombres (45.1%) que entre las mujeres (25.1%), estos hallazgos sugieren que el género podría desempeñar un papel relevante en las tasas de deserción, lo que indica la posible influencia de factores específicos que afectan de manera diferenciada a hombres y mujeres en su decisión de abandonar los estudios. Ver figura 6.

Figura 6

Distribución de las Variables Objetivos Después de Eliminar la Variable



Nota: Elaboración propia con base en estudios de Realinho et al., (2021).

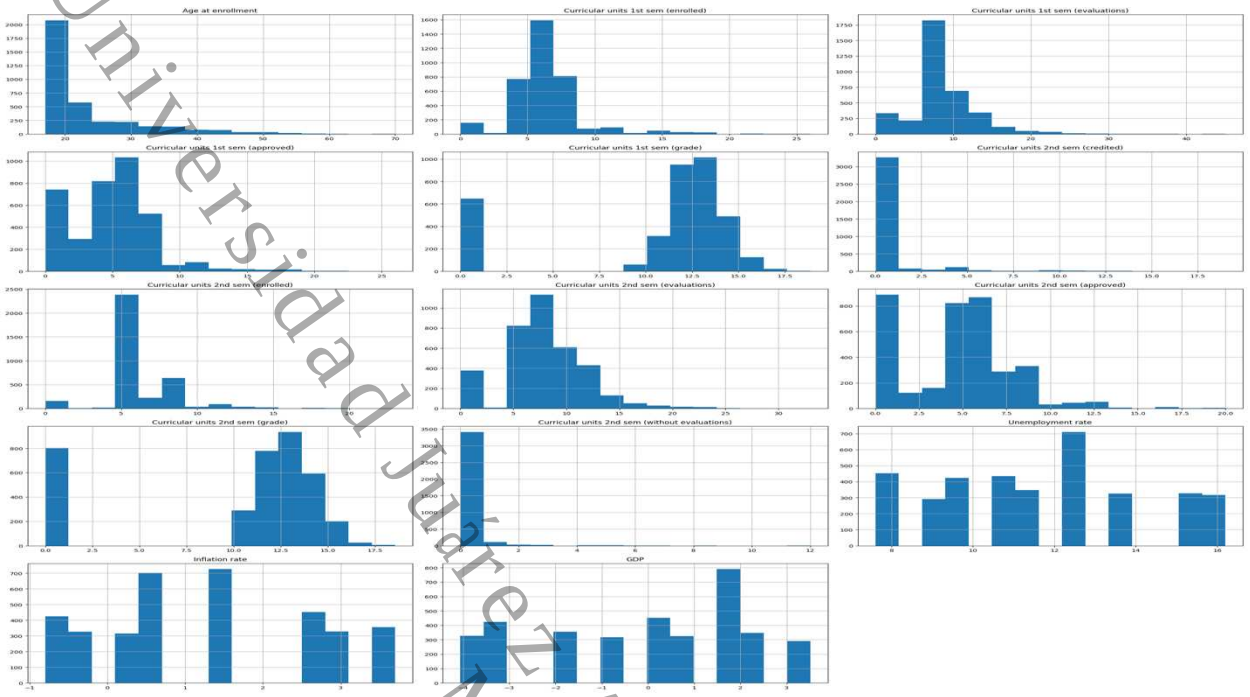
Pruebas de normalidad

Mediante el uso de histogramas, se visualizó la distribución de variables numéricas, como la edad de los estudiantes en el momento de la inscripción, el número de unidades curriculares en las que se matricularon y la cantidad de evaluaciones realizadas. El análisis de estas visualizaciones permitió identificar características clave de los datos, como la forma de la distribución, la presencia de valores atípicos y la concentración de valores en rangos específicos. Esta información resultó esencial para comprender la naturaleza del conjunto de datos y fundamentar decisiones en las etapas de preprocesamiento y análisis. Ver figura 7.

Por ejemplo, si un histograma revela una distribución con una cola larga hacia la derecha, esto podría indicar la presencia de valores atípicos que podrían influir negativamente en el análisis, en tales casos, se podría considerar aplicar técnicas de procesamiento de datos atípicos para mitigar el impacto de estos valores extremos.

Figura 7

Histogramas para visualizar la distribución de variables numéricas



Nota: Elaboración propia con base en estudios de Realinho et al., (2021).

Los histogramas se utilizan principalmente para representar la distribución de datos numéricos continuos, por lo que no resultan óptimos para visualizar variables nominales u ordinales. En la figura 8, se observa que solo 14 de las 35 características del conjunto de datos fueron graficadas mediante histogramas.

El análisis de los histogramas reveló que los datos no se distribuyen de manera normal. Por ejemplo, la distribución de la edad de los estudiantes al matricularse en la universidad muestra que la mayoría tienen entre 17 y 22 años, aunque también se identificaron estudiantes mayores de 35 años, posiblemente retomando sus estudios tras una pausa. En cuanto a las unidades curriculares, se observó que la mayoría de los estudiantes se matriculan y aprueban entre 5 y 10 unidades por semestre, con calificaciones medias cercanas a 3.0, factores económicos como el desempleo, la inflación y el PIB varían y podrían influir en el rendimiento académico y la estabilidad de los estudiantes.

Análisis de correlación

La correlación entre las variables es un aspecto clave durante el preprocesamiento de los datos, ya que permite identificar relaciones significativas que ayudan a comprender mejor el conjunto de datos. Además, este análisis facilita la detección de variables redundantes (Allah *et al.,* 2022). Para medir la correlación existen diversas técnicas, y en este caso se ha optado por utilizar la correlación de Spearman, dado que es menos susceptible a la influencia de valores atípicos y resulta más apropiada para variables que no siguen una distribución normal (El-Hashash y Shiekh, 2022).

En este trabajo se utilizó para identificar la relación entre las variables independientes y determinar si existía multicolinealidad, es decir, si algunas variables independientes estaban altamente correlacionadas entre sí.

La multicolinealidad puede suponer un problema en el aprendizaje automático, ya que dificulta la interpretación de los coeficientes del modelo y hace que este sea menos estable.

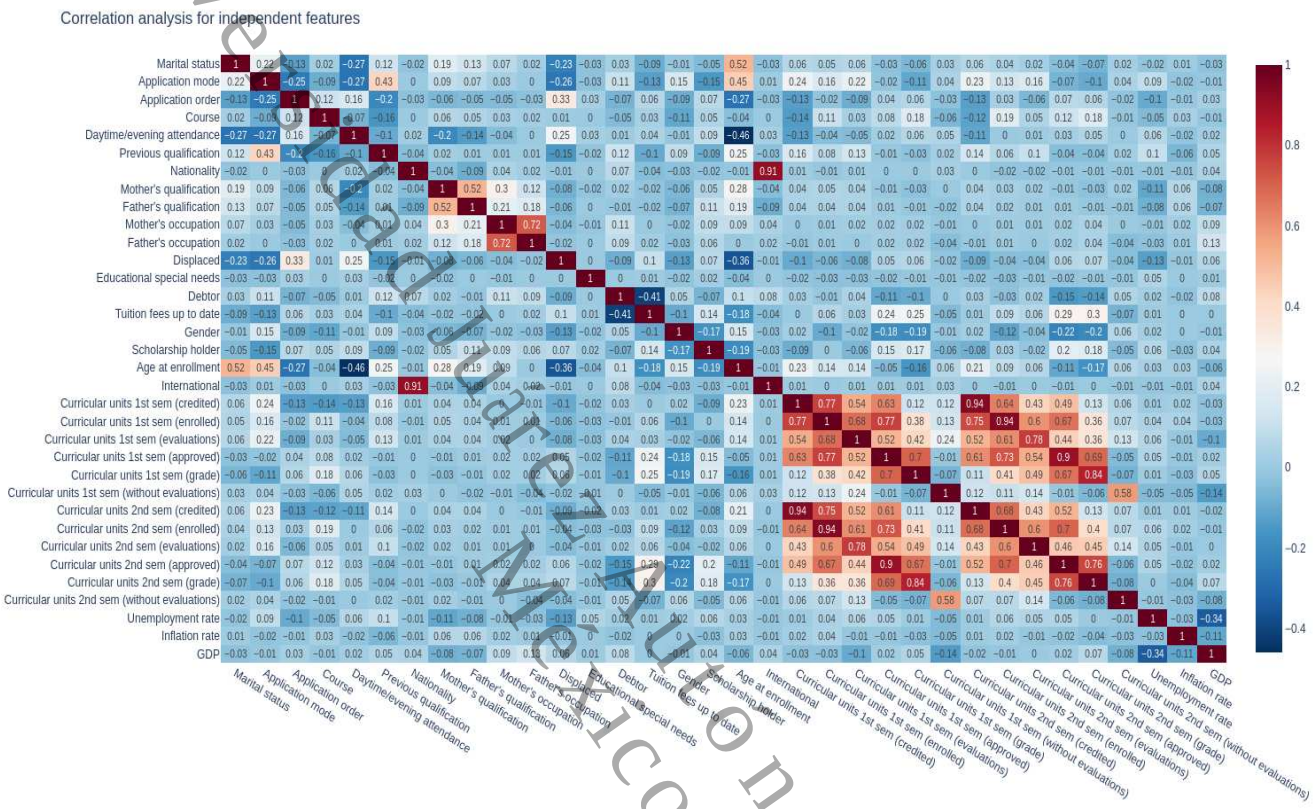
Para visualizar la correlación entre las variables, se utiliza un mapa de calor, este mapa muestra la correlación entre cada par de variables como un color, los colores más intensos indican una correlación más fuerte.

En este caso, el mapa de calor muestra que existen algunas variables con una alta correlación, como "Nacionalidad" e "Internacional", "Calificación de la madre" y "Calificación del padre", y "Ocupación de la madre" y "Ocupación del padre".

Ver figura 8.

Figura 8

Mapa de calor para identificar variables independientes correlacionadas entre si



Nota: Elaboración propia con base en estudios de Realinho et al., (2021).

Análisis de Componentes Principales

El Análisis de Componentes Principales es una técnica de reducción de dimensionalidad utilizada en el análisis de datos y el aprendizaje automático. Su objetivo principal es transformar un conjunto de variables correlacionadas en un nuevo conjunto de variables no correlacionadas, llamadas componentes principales. (de Zaballa & Carriegos, 2019)

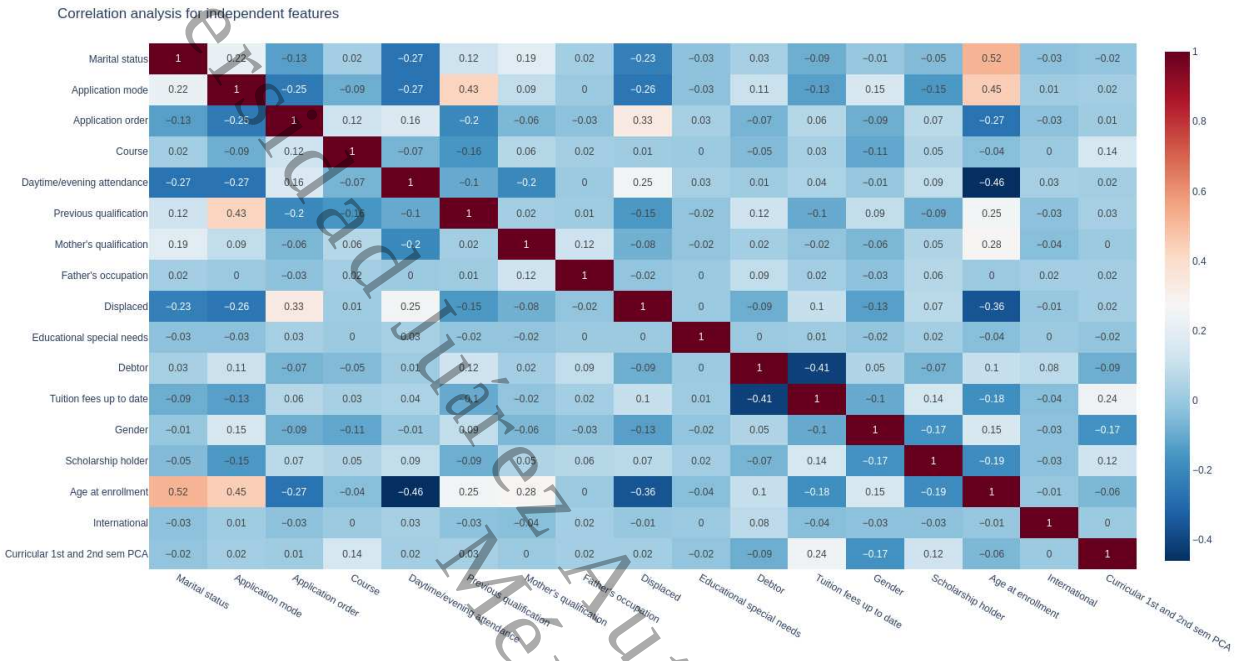
La observación de los resultados revela que ciertas características del conjunto de datos presentan una fuerte correlación entre sí.

Por ejemplo, las variables Nacionalidad e Internacional, así como las Calificaciones de la madre y las Calificaciones del padre, y las Ocupaciones de la madre y las Ocupaciones del padre están correlacionadas. Esto sugiere redundancia en los datos, lo que podría llevar a una duplicación de información innecesaria. Para simplificar el modelo, podríamos optar por eliminar una de cada pareja de variables correlacionadas, ya que ambas aportan información similar.

Sin embargo, en el caso de las variables relacionadas con las Unidades Curriculares del 1er semestre y las del 2do semestre, se observa una correlación significativa, pero ambas aportan información valiosa sobre el desempeño del estudiante en diferentes periodos. El riesgo aquí es que, al mantener ambas variables sin una gestión adecuada, podríamos inducir *overfitting* en el modelo debido a la redundancia de la información. Ver figura 9.

Figura 9

Mapa de calor después de eliminar variables independientes correlacionadas entre si



Nota: Elaboración propia con base en estudios de Realinho et al., (2021).

Análisis exploratorio de los datos

El Análisis Exploratorio de Datos (EDA), es un paso crucial en cualquier proyecto de análisis de datos o de aprendizaje automático. Su objetivo principal es entender la estructura, las características y las relaciones en los datos antes de aplicar modelos predictivos o realizar análisis más avanzados, esto permite descubrir patrones, detectar anomalías, probar hipótesis y verificar supuestos a través de estadísticas descriptivas y herramientas de visualización. En la tabla 5 se describen los estadísticos descriptivos de las variables cuantitativas del conjunto de datos.

Tabla 5

Tabla de medidas de tendencia central y dispersión

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
age at enrollment	3630	23.461	7.828	17	19	20	25	70
Curricular unit's 1st sem (enrolled)	3630	6.337	2.571	0	5	6	7	26
Curricular unit's 1st sem (evaluations)	3630	8.071	4.287	0	6	8	10	45
Curricular unit's 1st sem (approved)	3630	4.791	3.238	0	3	5	6	26
Curricular unit's 1st sem (grade)	3630	10.535	5.058	0	11	12.34	13.5	18.875
Curricular unit's 2nd sem (credited)	3630	0.582	2.023	0	0	0	0	19
Curricular unit's 2nd sem (enrolled)	3630	6.296	2.263	0	5	6	7	23
Curricular unit's 2nd sem (evaluations)	3630	7.763	3.964	0	6	8	10	33
Curricular unit's 2nd sem (approved)	3630	4.518	3.162	0	2	5	6	20
Curricular unit's 2nd sem (grade)	3630	10.036	5.482	0	10.518	12.333	13.5	18.571
Curricular unit's 2nd sem (without evaluations)	3630	0.142	0.748	0	0	0	0	12
Unemployment rate	3630	11.630	2.668	7.6	9.4	11.1	13.9	16.2
Inflation rate	3630	1.232	1.385	-0.8	0.3	1.4	2.6	3.7
Gdp	3630	-0.009	2.260	-4.0	-1.700	0.320	1.790	3.510

Nota: Elaboración propia con base en estudios de Realinho et al., (2021).

Identificación de datos faltantes

Otra etapa importante en el análisis de datos es la etapa de identificación y manejo de datos faltantes, Asegurarse de que no haya valores faltantes en nuestro conjunto de datos no solo garantiza un óptimo análisis, sino que también simplifica el proceso de modelización al evitar la necesidad de técnicas adicionales para manejar datos incompletos. Ver tabla 6.

Tabla 6

Visualización para observar si hay datos faltantes en nuestro dataset.

VARIABLE	VALORES FALTANTES
Marital status	0
Application mode	0
Application order	0
Course	0
Daytime/evening attendance	0
Previous qualification	0
Nationality	0
Mother's qualification	0
Father's qualification	0
Mother's occupation	0
Father's occupation	0
Displaced	0
Educational special needs	0
Debtor	0
Tuition fees up to date	0
Gender	0
Scholarship holder	0
Age at enrollment	0
International	0
Curricular unit's 1st sem (credited)	0
Curricular unit's 1st sem (enrolled)	0
Curricular unit's 1st sem (evaluations)	0
Curricular unit's 1st sem (approved)	0
Curricular unit's 1st sem (grade)	0
Curricular unit's 1st sem (without evaluations)	0
Curricular unit's 2nd sem (credited)	0
Curricular unit's 2nd sem (enrolled)	0
Curricular unit's 2nd sem (evaluations)	0
Curricular unit's 2nd sem (approved)	0
Curricular unit's 2nd sem (grade)	0
Curricular unit's 2nd sem (without evaluations)	0
Unemployment rate	0
Inflation rate	0
Gdp	0

Nota: Elaboración propia con base en estudios de Realinho et al., (2021).

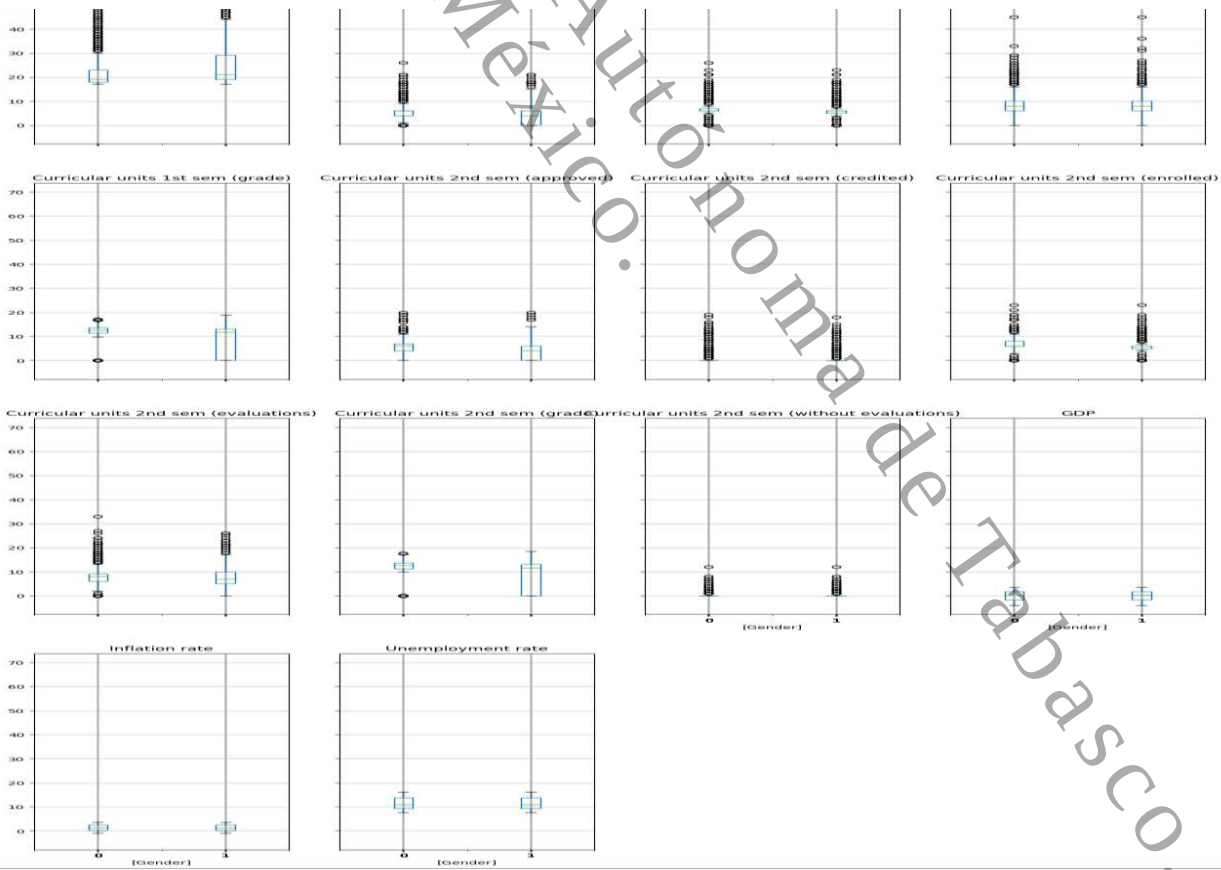
Identificación de datos atípicos

Un aspecto crucial del análisis exploratorio de datos es la detección de valores que se desvían significativamente del comportamiento típico o esperado, conocidos como datos

atípicos o outliers. Para identificar visualmente estas anomalías, se utilizó el diagrama de cajas y bigotes, que permite analizar qué características presentan este tipo de valores fuera de lo normal. Se pudo identificar que las variables que presentan datos atípicos incluyen: la edad de matrícula, las unidades curriculares aprobadas en el primer semestre, las unidades curriculares matriculadas en el primer semestre, las evaluaciones en el primer semestre, las calificaciones en el primer semestre, las unidades curriculares acreditadas en el segundo semestre, las unidades curriculares matriculadas en el segundo semestre, las evaluaciones en el segundo semestre, las unidades curriculares aprobadas en el segundo semestre, las calificaciones en el segundo semestre y las unidades curriculares sin evaluaciones en el segundo semestre. Ver figura 10.

Figura 10

Identificación de datos atípicos de variables numéricas



Nota: Elaboración propia con base en estudios de Realinho et al., (2021).

4.2. Preprocesamiento de datos

Winsorizacion

La winsorización es una técnica de preprocesamiento de datos que se utiliza para reducir el impacto de valores atípicos extremos en un conjunto de datos, en lugar de eliminar estos valores extremos, la winzorización los reemplaza con el valor más cercano dentro de un rango específico de percentiles. Este proceso ayuda a que luego los algoritmos los tome como datos normales dentro de un rango en específico, haciéndolos menos sensibles a estos valores extremos, lo cual puede mejorar la estabilidad y la precisión de los modelos al momento de entrenarlos.

Se identificó previamente que 11 de las 14 columnas numéricas contenían datos atípicos. En consecuencia, se aplicó la técnica correspondiente para tratarlos, después de este procedimiento, las columnas presentaron una distribución ajustada, como se observa en la figura 11.

Figura 11

Datos atípicos de las variables después de la winsorizacion.

Nota: Elaboración propia con base en estudios de Realinho et al., (2021).

Para el preprocesamiento de los datos numéricos, se aplicó la técnica RobustScaler con el objetivo de normalizar las características, minimizando así la influencia de los valores atípicos. Este método es especialmente útil para algoritmos como SVM y KNN, que son sensibles a las distancias entre puntos, la aplicación de esta técnica garantiza que las

variables con diferentes escalas no afecten de manera desproporcionada al rendimiento del modelo, lo que contribuye a una mayor estabilidad y precisión en los resultados.

4.3. Selección de características

El proceso de selección de características se llevó a cabo usando cinco técnicas distintas de selección. Se empleó un análisis de la varianza ANOVA para identificar 12 de las 35 características del conjunto de datos con las puntuaciones F más altas, posteriormente, se utilizó la técnica de la información mutua para identificar las 10 características con mayor dependencia estadística. Para las técnicas siguientes, el número de características seleccionadas se fijó en 12. La regresión logística, junto con un selector secuencial, se utilizó para identificar las variables que optimizan el rendimiento del modelo. Además, se emplearon las dos técnicas restantes, es decir, ANOVA y la técnica de información mutua, para identificar las 12 características más relevantes en cada caso. Tras la aplicación de estas técnicas, los archivos se guardaron en archivos CSV. De este modo se generaron cinco nuevos con juntos de datos, cada uno de ellos con las características seleccionadas por una técnica específica. Estos archivos se prepararon para el entrenamiento de los modelos. Ver Tabla 7.

Tabla 7

Características seleccionadas por cada una de las técnicas.

Técnica de selección de característica	Numero de características seleccionadas	Características seleccionadas
ANOVA	12	• Marital status • Application mode • Daytime/evening attendance • Previous qualification • Father's occupation • Displaced • Debtor • Tuition fees up to date • Gender • Scholarship holder • Age at enrollment • Curricular 1st and 2nd sem PCA

Mutual Information	10	• Father's occupation • Debtor • Tuition fees up to date • Gender • Scholarship holder • Age at enrollment • Curricular 1st and 2nd sem PCA
Sequential forward selection (SFS)	12	• Marital status, • Application mode • Daytime/evening attendance • Father's occupation • Displaced • Educational special needs • Debtor • Tuition fees up to date • Gender • Scholarship holder • International, • Curricular 1st and 2nd sem PCA
Recursive feature search (RFE)	12	• Marital • status • Application mode • Course, • Daytime/evening attendance • Father's occupation • Debtor • Tuition fees up to date • Gender • Scholarship holder • Age at enrollment • International, Curricular 1st and 2nd sem, PCA
LASSO	16	• Marital status, • Application mode, • Application order, • Course, • Daytime/evening attendance, • Previous qualification • Mother's qualification, • Father's occupation, • Displace,Educational • special needs, • Debtor,Tuition • fees up to date, • Gender, • Scholarship holder, • Age at enrollment, International, Curricular 1st and 2nd sem, PCA

Nota: Elaboración propia con base en estudios de Realinho et al., (2021).

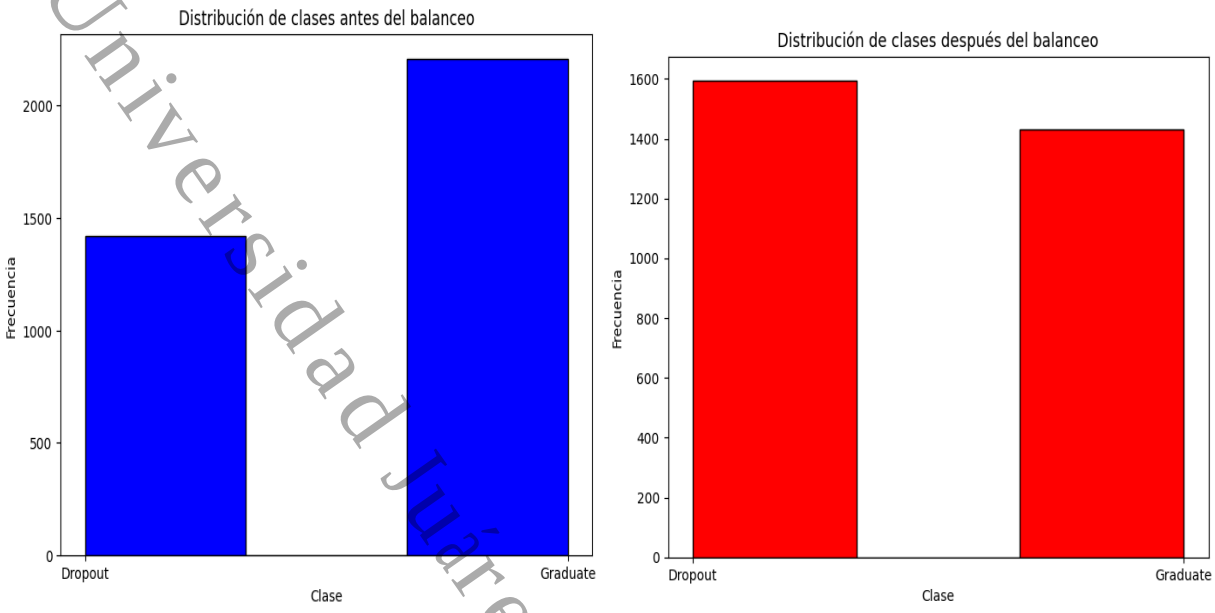
4.4. Entrenamiento

Balanceo de Datos

Para abordar el problema del desequilibrio de clases y mejorar el rendimiento de los modelos, se empleó la técnica SMOTEENN. Antes de su aplicación, el conjunto de datos mostraba un desequilibrio notable, con 2209 instancias de graduados y 1421 de abandonos. Tras la aplicación de SMOTEENN, el número de instancias de la clase de abandonos aumentó a 1657, mientras que el número de graduados se incrementó a 1602. Esto demuestra que SMOTEENN logró equilibrar las clases al generar nuevas instancias para la clase minoritaria (abandono), lo que permitió mejorar la capacidad del modelo para aprender de estos casos y ofrecer predicciones más precisas. Para mayor claridad, se presenta la distribución de las instancias en la Figura 12.

Figura 12

Antes y después de hacer balanceo de datos



Nota: Elaboración propia con base en estudios de Realinho et al., (2021).

Entrenamiento y prueba

Una vez procesado y equilibrado el conjunto de datos, se optó por una división 70-30, lo que implicó que el 70 % de los datos se utilizó para entrenar el modelo, mientras que el 30 % restante se destinó a probar el modelo ya entrenado. De este modo, fue posible entrenar el modelo con un conjunto de datos y evaluarlo con otro conjunto completamente diferente. De esta manera, se facilitó la evaluación del grado de generalización del modelo frente a datos desconocidos, lo que ayudó a prevenir el sobreajuste, que ocurre cuando el modelo presenta un buen desempeño en los datos de entrenamiento, pero un rendimiento deficiente en datos nuevos.

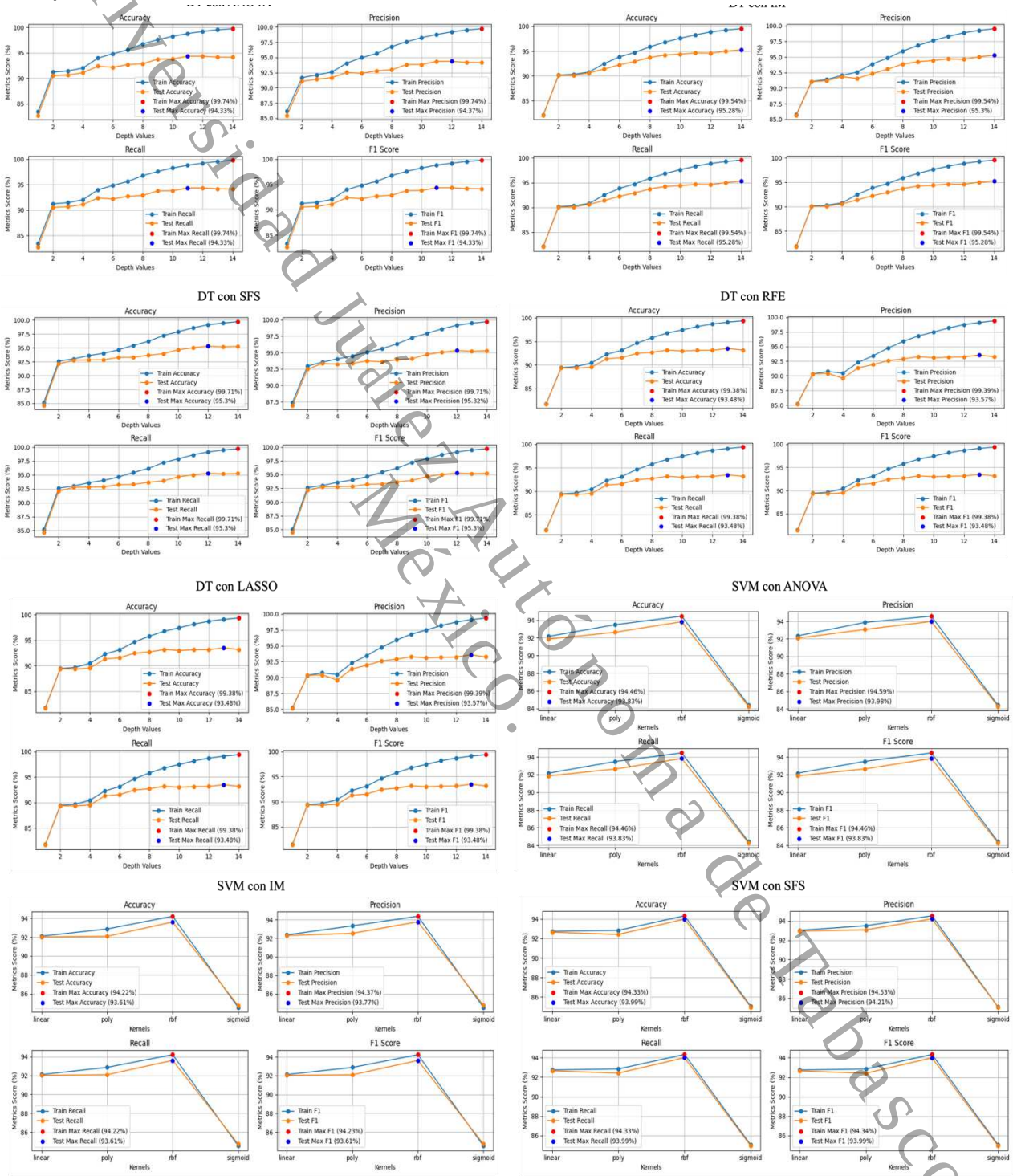
4.5. Evaluación y Comparación de los Modelos

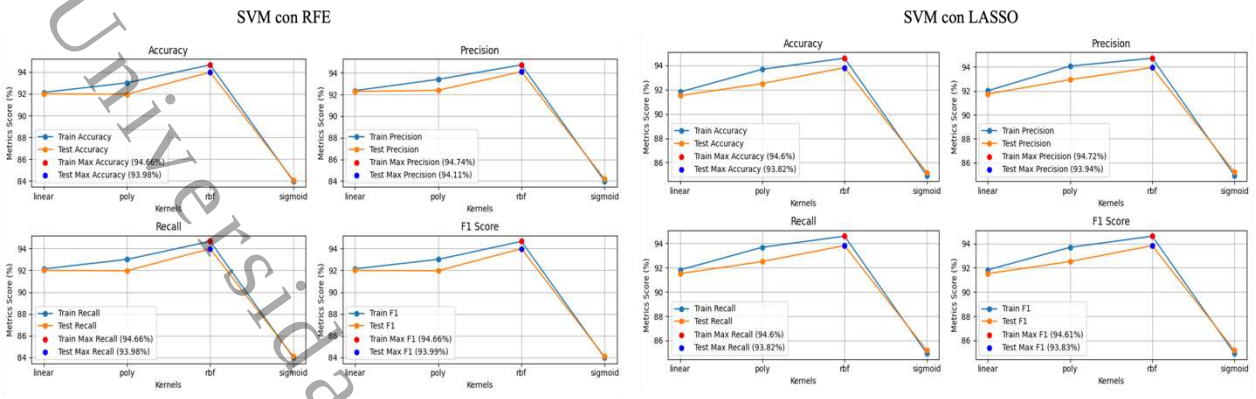
Después del entrenamiento, se aplicaron los algoritmos árboles de decisión, máquinas de vectores de soporte y KNN junto con cinco técnicas de selección de características: análisis de varianza, información mutua, regresión logística con selección secuencial (SFS), LASSO y eliminación recursiva de características (RFE).

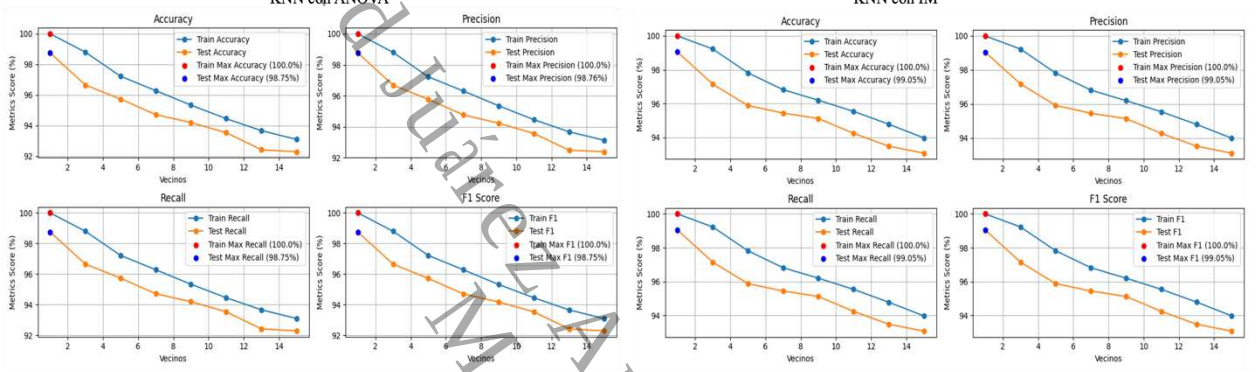
Los resultados obtenidos se analizaron comparando las métricas de precisión (accuracy), sensibilidad (recall) y puntuación F1 de cada combinación. Este análisis permitió identificar la combinación más efectiva para predecir la deserción escolar. Para consultar los detalles visuales de los resultados, consúltese la Figura 13.

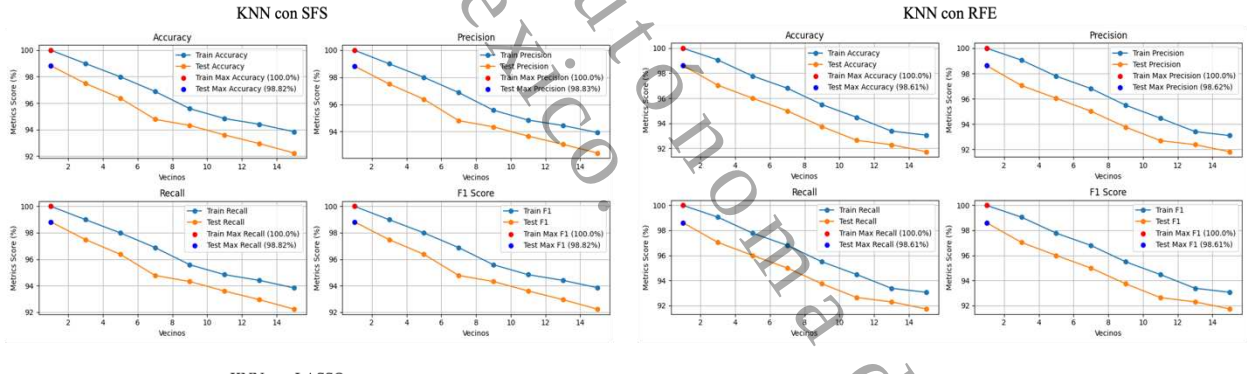
Figura 13

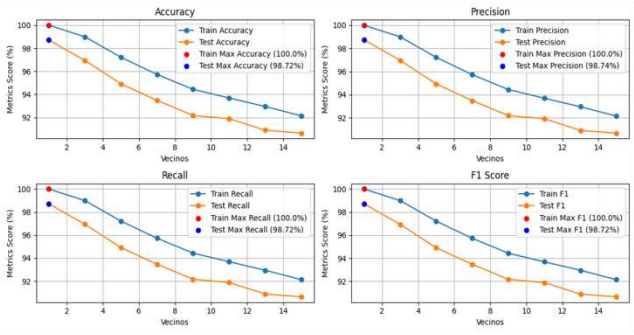
Comparación de las métricas de evaluación.











Nota: Elaboración propia

En la tabla 8 se muestran los resultados, expresados en porcentajes, de las métricas obtenidas para cada técnica de selección de características aplicada en combinación con los algoritmos utilizados, estos valores permiten evaluar el rendimiento de cada técnica en términos de precisión, exactitud, sensibilidad y puntuación F1.

Tabla 8

Resultados de métricas de clasificación

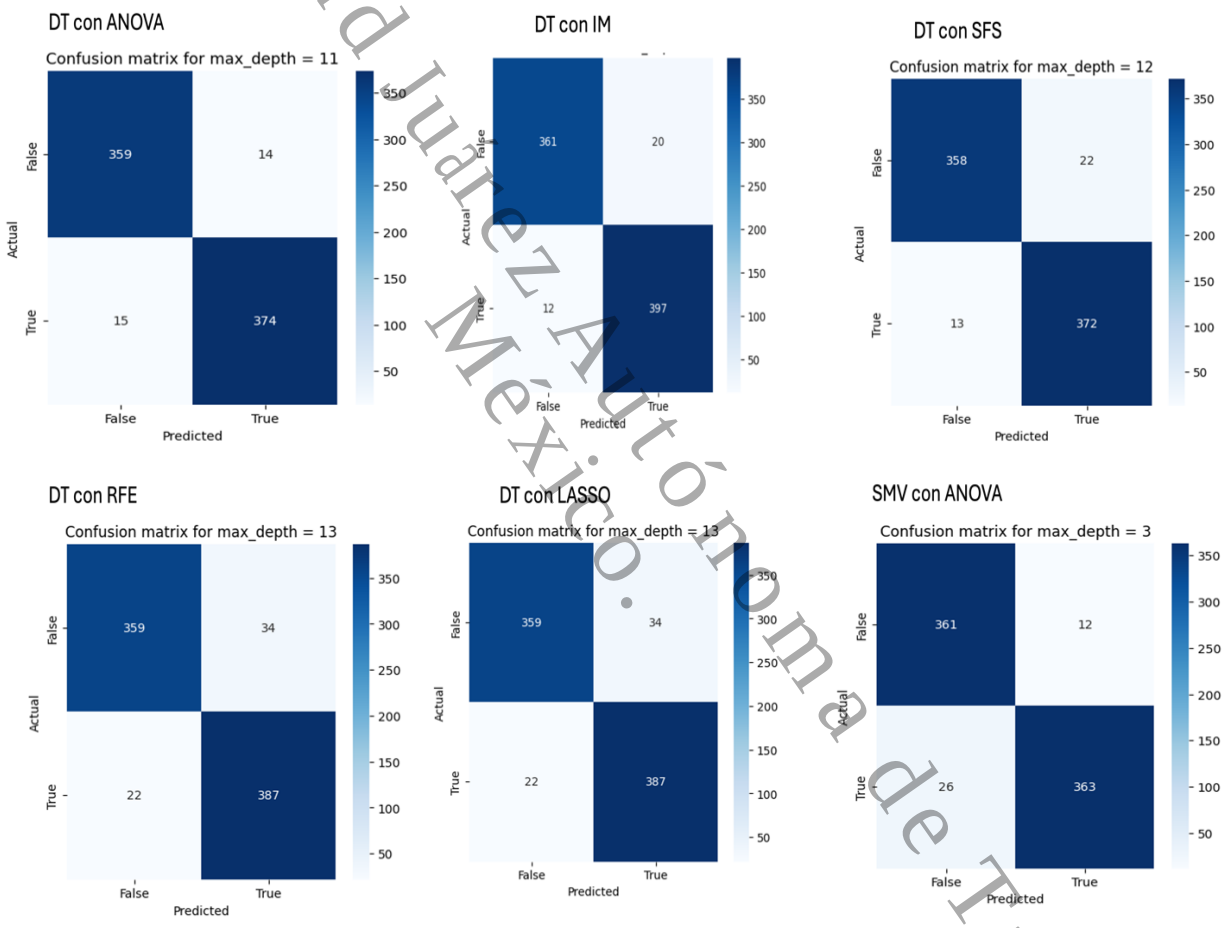
Algoritmos	Técnica de selección de característica	Precisión	Accuracy	Recall	F1 score
DT	ANOVA	94.37%	94.33%	94.33%	94.33%
	Mutual Information	95.3%	95.28%	95.28%	95.28%
	Sequential forward selection (SFS)	95.32%	94.3%	94.3%	94.3%
	Recursive feature search (RFE)	93.57%	94.48%	94.48%	94.48%
	LASSO	93.57%	94.48%	94.48%	94.48%
SVM	ANOVA	93.98%	93.83%	93.83%	93.83%
	Mutual Information	93.77%	93.61%	93.61%	93.61%
	Sequential forward selection (SFS)	94.21%	93.99%	93.99%	93.99%
	Recursive feature search (RFE)	94.11%	93.98%	93.98%	93.98%
	LASSO	93.94%	93.82%	93.82%	93.82%
KNN	ANOVA	98.76%	98.75%	98.75%	98.75%
	Mutual Information	99.05%	99.05%	99.05%	99.05%
	Sequential forward selection (SFS)	98.83%	98.82%	98.82%	98.82%
	Recursive feature search (RFE)	98.62%	98.61%	98.61%	98.61%
	LASSO	98.74%	98.72%	98.72%	98.72%

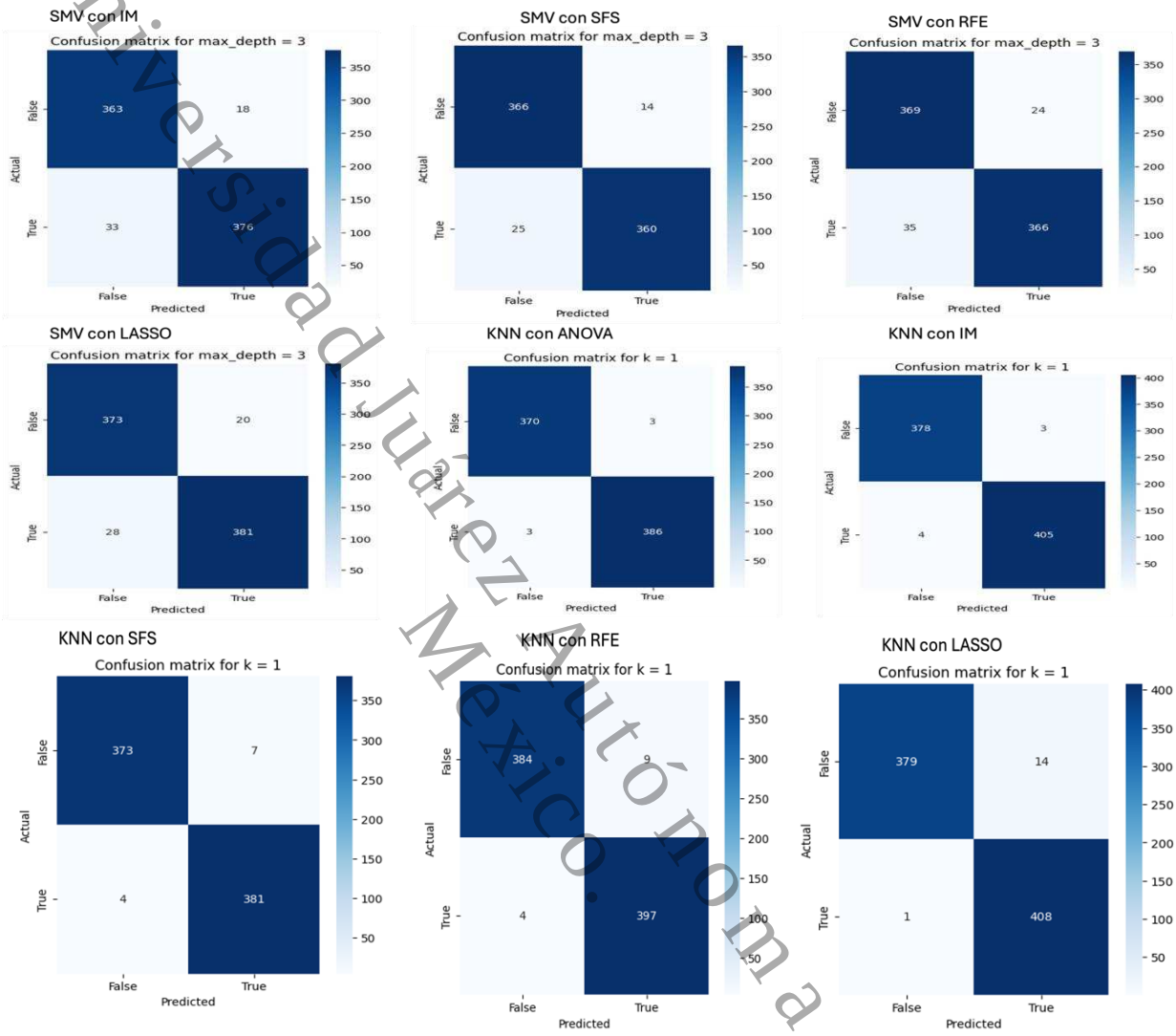
Nota: Elaboración propia.

En la Figura 14 muestra los valores de las matrices de confusión obtenidos para cada técnica de selección de características y algoritmos, detallando los aciertos y errores en las predicciones de cada modelo.

Figura 14

Comparación de resultados de las matrices de confusión de cada modelo entrenado





Nota: Elaboración propia

En la tabla 9 se puede observar las tasas de TN, FN, FP y TP, La matriz compara los resultados de las predicciones del modelo con los valores reales, muestra cuántas veces el modelo hizo predicciones correctas o incorrectas.

Tabla 9

Tasas de resultados de la matriz de confusión por cada selección de características.

Algoritmos	Técnica de selección de características	TN	FN	FP	TP
DT	ANOVA	359	14	15	374
	Mutual Information	361	20	12	397
	Sequential forward selection (SFS)	358	22	13	372
	Recursive feature search (RFE)	359	34	22	387
	LASSO	359	34	22	387
SVM	ANOVA	361	12	26	363
	Mutual Information	363	18	33	376
	Sequential forward selection (SFS)	366	14	25	360
	Recursive feature search (RFE)	369	24	35	366
	LASSO	377	20	28	381
KNN	ANOVA	370	3	3	386
	Mutual Information	378	3	4	405
	Sequential forward selection (SFS)	373	7	4	381
	Recursive feature search (RFE)	384	9	4	397
	LASSO	379	1	14	408

Nota: Elaboración propia

4.6. Mejora de los modelos de predicción del abandono mediante técnicas de selección de características

En este estudio se entrenaron y evaluaron 15 modelos mediante validación cruzada esta fue una de las principales garantías de la solidez de nuestros resultados fue la implementación de la validación cruzada de 10, es el número base estándar que se utiliza generalmente. Esto nos permitió evaluar la consistencia y la fiabilidad de los modelos,

esto ayudó a reducir el riesgo de sobreajuste *(overfitting)* y aseguró que los modelos pudieran generalizar correctamente nuevos datos.

Los modelos obtenidos en este estudio se generaron combinando tres algoritmos de clasificación de aprendizaje automático y cinco técnicas diferentes de selección de características. Dado el contexto particular del estudio y las características del conjunto de datos, se prestó especial atención a la métrica de precisión.

La precisión es una métrica clave que evalúa la fiabilidad de las predicciones positivas realizadas por el modelo. En otras palabras, mide cuán acertadas son las predicciones del modelo cuando clasifica una instancia como positiva. Un modelo con alta precisión indica que tiene una baja tasa de errores al etiquetar casos como positivos, lo que sugiere que la mayoría de las instancias clasificadas como positivas son efectivamente correctas. Esta métrica es crucial en contextos como el presente estudio, donde las intervenciones basadas en predicciones erróneas pueden tener consecuencias significativas.

Algoritmos de clasificación usados

Se seleccionaron los algoritmos KNN, árboles de decisión y SVM debido a su combinación de simplicidad, interpretabilidad y capacidad predictiva, características esenciales para el análisis de predicción. El algoritmo KNN clasifica a los estudiantes en función de las similitudes directas con otros casos, lo que lo convierte en una opción intuitiva y fácil de implementar, por su parte, los árboles de decisión proporcionan una representación visual clara del proceso de toma de decisiones, lo que facilita su comprensión y aplicación por parte de las instituciones educativas, finalmente, la SVM se distingue por su efectividad a la hora de separar datos complejos y no lineales, lo que asegura una clasificación precisa, incluso cuando los factores asociados a la deserción resultan difíciles de distinguir.

Técnica de selección de características

Se optó por las técnicas de selección de características Lasso, ANOVA, Mutual Information (MI), Sequential Forward Selection (SFS) y Recursive Feature Elimination (RFE) debido a su capacidad para optimizar el rendimiento del modelo al reducir la cantidad de características irrelevantes o redundantes, lo cual mejora tanto la precisión como la eficiencia. Lasso permite eliminar automáticamente las características menos significativas, lo que contribuye a simplificar el modelo. Por otro lado, ANOVA y MI son útiles para identificar las características que tienen un impacto significativo, especialmente en situaciones en las que las relaciones no son evidentes o son no lineales. SFS y RFE seleccionan las mejores características de manera secuencial y recursiva, asegurando que solo se mantengan los factores más relevantes para predecir la deserción, y se identificó que, en común, las cinco técnicas de selección de características priorizan las variables sociodemográficas, las cuales desempeñan un papel crucial en la predicción del abandono escolar.

Configuración de los Hiperparámetros

Como parte de la configuración experimental para el proceso de entrenamiento con las cinco técnicas de selección de características, se ajustaron los hiperparámetros de los algoritmos. En el caso del algoritmo de árbol de decisión, se evaluaron distintos valores del parámetro max_depth, con un rango comprendido entre 1 y 14. El objetivo del ajuste fue encontrar un equilibrio adecuado entre la simplicidad del modelo y su capacidad de ajuste para evitar tanto el sobreajuste (overfitting) como el subajuste (underfitting), garantizando así un desempeño óptimo del modelo en términos de generalización y precisión.

Se evaluaron las máquinas de vectores de soporte (SVM) utilizando diferentes tipos de kernel: lineal, polinómico (poli), de función de base radial (RBF) y sigmoide. La selección de estos cuatro kernels permitió determinar cuál resultaba más adecuado para el conjunto de datos, ya que no siempre es evidente de antemano si los datos son linealmente separables o requieren transformaciones más complejas.

Por otro lado, el modelo KNN se entrenó utilizando valores impares para el parámetro k, como 1, 3, 5, 7, 9, 11, 13 y 15, ya que es una práctica común emplear números impares para evitar empates en el proceso de votación y garantizar una decisión clara. El objetivo de este ajuste fue encontrar el valor óptimo de k que ofreciera la mejor precisión en las predicciones de clasificación.

Resultados Detallados por Modelo y Técnica

Árboles de Decisión (DT)

La técnica de Selección Secuencial Adelante (SFS) arrojó los mejores resultados para este modelo, alcanzando las siguientes métricas:

- **Precisión (Accuracy):** 95.33%
- **Exactitud (Precision):** 94.30%
- **Sensibilidad (Recall):** 94.30%
- **F1-score:** 94.30%

Estos resultados indican que el modelo basado en SFS con Árboles de Decisión es eficiente en la clasificación de los estudiantes, aunque presenta un ligero margen de mejora en comparación con KNN.

Máquinas de Vectores de Soporte (SVM)

El modelo SVM mostró un comportamiento óptimo cuando se utilizó la técnica de Selección Secuencial Adelante (SFS), obteniendo los siguientes resultados:

- **Precisión (Accuracy):** 94.21%
- **Exactitud (Precision):** 93.99%
- **Sensibilidad (Recall):** 93.99%
- **F1-score:** 93.99%

El rendimiento del SVM es similar al de los Árboles de Decisión cuando se utiliza SFS, demostrando ser una técnica efectiva para este modelo. Sin embargo, aún queda por debajo de KNN en términos de precisión global.

K-Vecinos más Cercanos (KNN)

El modelo KNN alcanzó los mejores resultados generales cuando se utilizó la técnica de Información Mutua (MI), destacando en todas las métricas evaluadas:

- **Precisión (Accuracy):** 99.05%
- **Exactitud (Precision):** 99.05%
- **Sensibilidad (Recall):** 99.05%
- **F1-score:** 99.05%

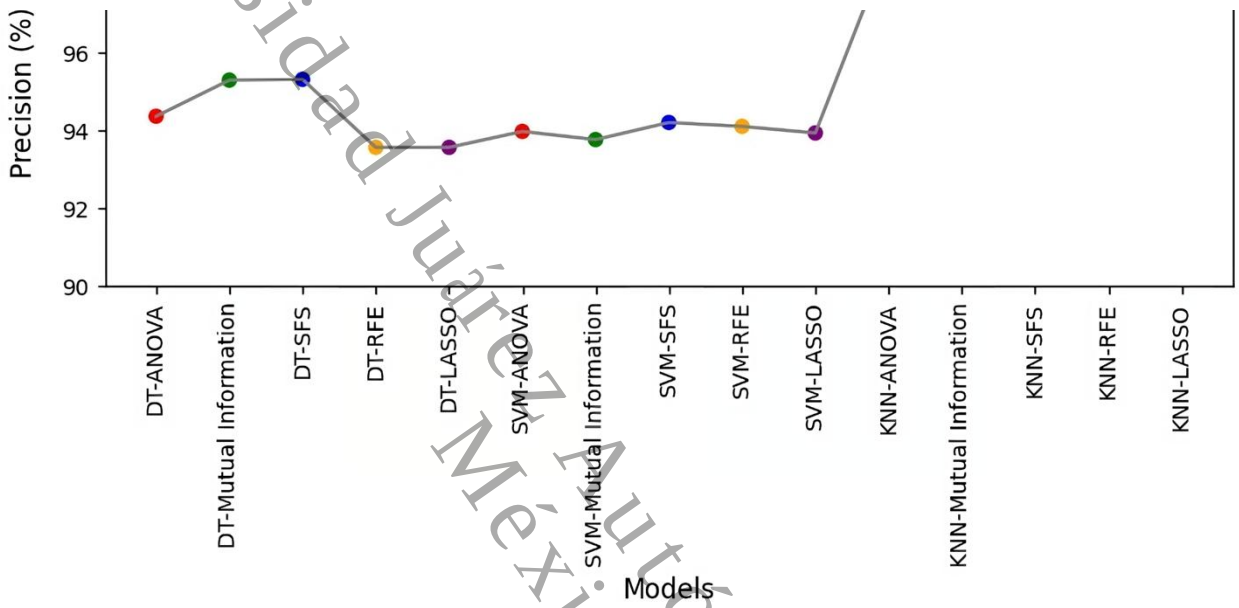
Este rendimiento sugiere que el modelo KNN, en combinación con la técnica MI, es altamente eficaz para predecir el riesgo de deserción escolar, superando significativamente a los otros algoritmos evaluados.

La Figura 15, muestra los resultados de las métricas de precisión en la fase de prueba,

de las 5 técnicas de selección de características entrenadas por los 3 algoritmos de clasificación.

Figura 15

Resultados de la métricas precisión



Nota: Elaboración propia

Estos resultados sugieren que el modelo KNN, utilizando la técnica MI, ofrece el mejor rendimiento predictivo, ya que obtuvo las puntuaciones más altas en todas las métricas. Además, la técnica SFS demostró resultados superiores para los algoritmos DT y SVM, superando a otros métodos de selección de características.

4.7. Comparación con el estado del arte

Al comparar los resultados de este estudio con los hallazgos previos publicados en la literatura, es posible evaluar la efectividad de las técnicas de selección de características y de los modelos de aprendizaje automático implementados para predecir la deserción escolar. Investigaciones anteriores han identificado factores clave, como las variables socioeconómicas y el rendimiento académico temprano, como determinantes importantes en el proceso de deserción. Sin embargo, cada estudio presenta enfoques metodológicos y contextuales distintos, lo que puede influir en la precisión de los modelos y en la relevancia de las características seleccionadas. La Tabla 10 presenta una comparación entre los modelos generados en esta investigación

Tabla 10

Comparación de estudios relevantes de modelos predictivos de deserción escolar usando técnicas de selección de características y resultados de esta tesis.

TRABAJO	ALGORITMOS DE APRENDIZAJE AUTOMÁTICO	TÉCNICA DE SELECCIÓN DE CARACTERÍSTICAS	RESULTADOS
Youssef, m *et al.,* (2019)	• SVM, • Decision Tree • Naive Bayes • LR • KNN	Sequential Forward Selection, Sequential Backward Selection, Recursive Feature Elimination	After SVM AUC 0.972 Precision 0.988 Recall 0.980 F1-score 0.984 Before RF SFS AUC 0.962 Precision 0.989 Recall 0.964 F1-score 0.976
Zapata-medina *et al.,* (2024)	• SVM • RF • GBT	FSmRMR, FS Boruta FS LASSO,	After RF Precision 0.920 Recall 0.570 F1-score 0.700

		FS GA, FS PSO	Before FS LASSO GBT Precision 0.924 Recall 0.632 F1-score 0.751
Li *et al.*, (2018)	• LR • SVM • RF	Mutual information (MI), random forest (RF), and recursive feature elimination (RFE)	After LR AUC 0.8630 Precision 0.8567 Recall 0.8629 F1-score 0.8449 Before LR RF AUC 0.8629 Precision 0.8573 Recall 0.8633 F1-score 0.8472
TESIS, 2024	DT KNN SMV	ANOVA MI LASSO SFS RFE	Precision 0.9905 Accuracy 0.9825 Recall. 0.9882 F1-score. 0.9861

Nota: Elaboración Propia

En general, los resultados de este estudio superan los de la literatura en cuanto a las métricas de precisión, recall y F1-score, y destacan particularmente la capacidad predictiva del modelo KNN cuando se emplea la técnica de información mutua.

Estos resultados indican que las técnicas de selección de características y los algoritmos empleados en esta investigación son más eficaces para identificar a los estudiantes con riesgo de deserción escolar.

Este desempeño podría tener implicaciones significativas en el diseño de modelos de intervención educativa temprana y en la optimización de recursos destinados a la prevención de la deserción escolar universitaria.

Capítulo V. Conclusiones y Recomendaciones

En este trabajo, se evaluaron distintos algoritmos de clasificación combinados con técnicas de selección de características para obtener el mejor modelo que pudiera predecir la deserción escolar. Se utilizaron árboles de decisión, máquinas de vectores de soporte y vecinos más cercanos

El análisis de los modelos de clasificación reveló resultados relevantes sobre la implementación de técnicas de selección de características. En particular, la combinación de árboles de decisión (DT) con la técnica de selección secuencial adelante (SFS) permitió alcanzar una precisión del 95,32 %. Además, el desempeño en métricas como la exactitud, la sensibilidad y la puntuación F1 fue consistente, con valores del 94,30 % en cada caso. Estos resultados evidencian que el modelo logró identificar correctamente una alta proporción de instancias, manteniendo un equilibrio adecuado entre verdaderos positivos y estabilidad en el desempeño general.

De manera complementaria, las Máquinas de Vectores de Soporte (SVM), con SFS, alcanzaron una precisión del 94.21%, mientras que la *accuracy*, *recall* y *F1-Score* se mantuvieron en 94.30%. Si bien su rendimiento fue ligeramente inferior al de los Árboles de Decisión, este modelo mostró un alto grado de consistencia y un buen equilibrio en la clasificación de instancias, lo que lo coloca en un rango competitivo en términos de predicciones correctas.

El modelo basado en el algoritmo KNN, combinado con la técnica de Información Mutua para la selección de características, obtuvo resultados excepcionales. Alcanzó una precisión, exactitud, recall y F1-score del 99,05 % en cada métrica, estos resultados evidenciaron que el algoritmo KNN, al integrar esta técnica, mostró una eficacia notable a la hora de realizar predicciones precisas, superando de manera significativa el desempeño de los otros modelos evaluados en todas las métricas consideradas.

De manera general la técnica SFS demostró ser la más eficaz para los algoritmos de Árboles de Decisión y SVM, seleccionando el subconjunto óptimo de características para maximizar su rendimiento, sin embargo, el uso la tecnica de Información Mutua con KNN destacó particularmente, ya que permitió que este modelo alcanzara la precisión de todos los evaluados y con la gran mayoría de las predicciones correctas, además, la superioridad del modelo KNN también se confirmó mediante la matriz de confusión, donde mostró un bajo porcentaje de falsos negativos (FN) y un alto número de verdaderos positivos (TP), lo que lo posiciona como el mejor modelo en términos de desempeño general.

Cabe destacar que todos los métodos de selección de características identificaron factores sociodemográficos, como el estado civil, la ocupación del padre y la edad y la edad en el momento de la matriculación, estas variables desempeñan un papel que influyen en los resultados académicos y deberían ser prioritarias y consideradas en futuras creaciones de conjuntos de datos y posteriormente en modelos de predicción del abandono.

En futuras investigaciones, se plantea la posibilidad de integrar algoritmos de clasificación avanzados, como las redes neuronales profundas, así como enfoques bioinspirados, como los algoritmos genéticos o de optimización por enjambre. Estas técnicas no solo podrían incrementar la capacidad predictiva de los modelos, sino también abordar problemas complejos de manera más eficaz. Además, se sugiere utilizar métodos especializados para seleccionar características de variables no numéricas, como codificaciones avanzadas o análisis de componentes categóricos, lo que permitiría incluir información cualitativa importante en los modelos.

Por otra parte, explorar algoritmos diseñados para identificar características relevantes podría ampliar la capacidad de los modelos para capturar las relaciones subyacentes en los datos. Este enfoque no solo mejoraría la precisión y la generalización de las

predicciones, sino que también proporcionaría una comprensión más detallada de los factores asociados a la deserción escolar. Este conocimiento sería fundamental para desarrollar intervenciones educativas más focalizadas y efectivas, optimizando recursos y maximizando el impacto en la retención estudiantil. Estas líneas de investigación ofrecen un camino prometedor para avanzar en el campo del aprendizaje automático aplicado a problemas educativos.

Referencias citadas

Abarca, A., y Sánchez, M. (2005). La deserción estudiantil en la educación superior: El caso de la Universidad de Costa Rica. *Revista Actualidades Investigativas en Educación, 5*(Número). https://www.redalyc.org/articulo.oa?id=44759911

Adelman, C. (1999). *Answers in the toolbox: Academic intensity, attendance patterns, and bachelor's degree attainment.* U.S. Department of Education. https://eric.ed.gov/?id=ED431363

Agrawal, H., y Mavani, H. (2015). Student performance prediction using machine l earning. *International Journal of Engineering Research, 4*(3), 111–113. https://doi.org/10.17577/ijertv4is030127

Alayo, F. (2020, julio 13). Unos 174.000 estudiantes peruanos dejaron la universidad en lo que va del 2020. *El Comercio.* https://elcomercio.pe/lima/sucesos/unos-174000-estudiantes-peruanos-dejaron-la-universidad-en-lo-que-va-del-2020-noticia/

Alonso-Betanzos, A. (2007). Filter methods for feature selection: A comparative study. En *Proceedings of the International Conference on Intelligent Data Engineering and Automated Learning* (pp. 178–187). Birmingham, UK. https://dl.acm.org/doi/10.1007/978-3-540-77226-2_19

Amartya, S., y Kundan, K. D. (2007). *Application of Data mining Techniques in Bioinformatics* (Tesis de Ingeniería en Ciencias de la Computación). National Institute of Technology, (Deemed University), Rourkela. http://ethesis.nitrkl.ac.in/4154/

Amin, A., Takib, R., Raza, S., & Javed, S. (2014). Extract association rules to minimize the effects of dengue by using a text mining technique. *Journal of Infectious Diseases, 3*(4).

https://www.researchgate.net/publication/320609849_EXTRACT_ASSOCIATION_RULES_TO_MINIMIZE_THE_EFFECTS_OF_DENGUE_BY_USING_A_TEXT_MINING_TECHNIQUE

Ander-Egg, E. (2006). Métodos y Técnicas de investigación. *Social III, 3*, 13–36. https://epiprimero.wordpress.com/wp-content/uploads/2012/01/ander-egg-tecnicas-de-investigacion-social.pdf

Aquino, A. A., Molero, G., & Rojano, R. (2015). Hacia un nuevo proceso de minería de datos centrado en el usuario. *Pistas Educativas*, 114. Instituto Tecnológico de Celaya. https://pistaseducativas.celaya.tecnm.mx/index.php/pistas/article/view/303/0

Arbeláez-Campillo, D. F., Villasmil, J. J., & Rojas-Bahamón, M. J. (2021). Inteligencia artificial y condición humana: ¿Entidades contrapuestas o fuerzas complementarias? *Revista de Ciencias Sociales (Ve), XXVII*(2), 502–513. https://doi.org/10.31876/rcs.v27i2.35937

Báez, I. H. (2016). Clasificador Bayesiano Ingenuo en RapidMiner [Tesis de grado, Benemérita Universidad Autónoma de Puebla]. Repositorio Institucional BUAP. http://repositorioBUAP.com/handle/123456789/7695

Baker, R., & Yacef, K. (2009). El estado de la minería de datos educativos en 2009: A Review and Future Visions. *Journal of Educational Data Mining, 1*(1), 3–17. https://jedm.educationaldatamining.org/index.php/JEDM/article/view/8

Ballesteros, E., & Benalcázar, A. (2017). Principales factores que inciden en la deserción de estudiantes de la carrera de Ingeniería Comercial de la Universidad de Guayaquil durante el período 2011–2015 [Tesis de pregrado, Universidad de Guayaquil]. Repositorio Universidad de Guayaquil. http://repositorio.ug.edu.ec/handle/redug/20581

Barrientos, M. R. C. (2009). Árboles de decisión como herramienta en el diagnóstico médico. *Revista Médica de la Universidad Veracruzana, 9*(2), 19–24. https://www.medigraphic.com/pdfs/veracruzana/muv-2009/muv092c.pdf

Berens, J., Schneider, K., Görtz, S., Oster, S., & Burghoff, J. (2019). Early detection of students at risk - Predicting student dropouts using administrative student data and machine learning methods. *Journal of Educational Data Mining, 11*(3), 1–41. https://doi.org/10.5281/zenodo.3594771

Brezočnik, L., Nalli, G., De Leone, R., Val, S., Podgorelec, V., & Karakatič, S. (2023, mayo). Machine Learning Model for Student Drop-Out Prediction Based on Student Engagement. En *International Conference "New Technologies, Development and Applications"* (pp. 486–496). Cham: Springer Nature Switzerland. https://repository.mdx.ac.uk/item/wx656

Brito Sarasa, R., Rosete Suárez, A., & Acosta Sánchez, R. (2018). Desarrollo de un proceso de KDD en el ámbito docente: Preparación de los datos. *CUAJAE*, 2–7. https://tesla.puertomaderoeditorial.com.ar/index.php/tesla/article/view/226

Buendía, L., & Colás, M. (1999). F. Hernández (2005). *Metodología de la Investigación.* México: Ediciones Hill Interamericana. Constitución de la República Bolivariana de Venezuela. https://josedominguezblog.wordpress.com/wp-content/uploads/2015/06/investigacion-fundamentos-y-metodologia.pdf

Cabrera, A. F., Nora, A., & Castañeda, M. B. (1992). The role of finances in the persistence process: A structural model. *Research in Higher Education, 33*(5), 303–336. https://link.springer.com/article/10.1007/BF00973759

Cabrera, A. F., Nora, A., & Castañeda, M. B. (1993). College persistence: Structural equations modelling test of integrated model of student retention. *Journal of Higher Education, 64*(2), 123–320. https://eric.ed.gov/?id=EJ461421

Camargo García, A. J. (2020). *Modelo para la predicción de la deserción de estudiantes de pregrado, basado en técnicas de minería de datos* (Tesis doctoral). Corporación Universidad de la Costa. https://repositorio.cuc.edu.co/entities/publication/f9b9657c-4873-4193-96bb-b502a7fe3962

Canales, A., & De los Ríos, D. (2007). Factores explicativos de la deserción universitaria. *Calidad en la Educación, 26*, 173–201. https://doi.org/10.31619/caledu.n26.239

Castaño, E., Gallón, S., Gómez, K., & Vásquez, J. (2008). Análisis de los factores asociados a la deserción estudiantil en la educación superior: Un estudio de caso. *Revista de Educación, 345*, 255–280. https://dialnet.unirioja.es/servlet/articulo?codigo=2506107

Castrillón, O. D., Sarache, W., & Ruiz-Herrera, S. (2020). Prediction of academic performance using artificial intelligence techniques. *Formación Universitaria, 13*(1), 93–102. https://doi.org/10.4067/S0718-50062020000100093

Centro de Microdatos. (2008). *Estudios sobre las causas de deserción universitaria: Informe final.* Chile. https://www.opech.cl/educsuperior/politica_acceso/informe_final_causas_desercion_universitaria.pdf

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://dl.acm.org/doi/10.1145/2939672.2939785

Clow, D. (2013). An overview of learning analytics. *Teaching in Higher Education, 18*(6), 683–695. https://www.tandfonline.com/doi/abs/10.1080/13562517.2013.827653

Contreras, L. E., Fuentes, H. J., & Rodríguez, I. (2020). Academic interruption model using automatic learning algorithms. *International Journal of Mechanical and Production*

Engineering Research and Development (IJMPERD), 10(3), 16075–16086. http://www.tjprc.org/publishpapers/2-67-1602700574-IJMPERDJUN20201525.pdf

Cruz, E., González, M., & Rangel, J. (2022). Técnicas de machine learning aplicadas a la evaluación del rendimiento y a la predicción de la deserción de estudiantes universitarios, una revisión. *Prisma Tecnológico, 13*(1), 77–87. https://doi.org/10.33412/pri.v13.1.3039

Cuevas-Chávez, P. A., Narciso, S., Sánchez-Jiménez, E., Pérez, I. C., Hernández, Y., & Ortiz-Hernández, J. (2024). School dropout prediction with class balancing and hyperparameter configuration. En H. Calvo, L. Martínez-Villaseñor, H. Ponce, R. Zatarain Cabada, M. Montes Rivera, & E. Mezura-Montes (Eds.), *Advances in Computational Intelligence. MICAI 2023 International Workshops. MICAI 2023. Lecture Notes in Computer Science* (Vol. 14502). Springer, Cham. https://doi.org/10.1007/978-3-031-51940-6_2

Dake, D. K., & Buabeng-Andoh, C. (2022). Using Machine Learning Techniques to Predict Learner Drop-out Rate in Higher Educational Institutions. *Mobile Information Systems, 2022*(1), 2670562. https://onlinelibrary.wiley.com/doi/10.1155/2022/2670562

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems, 49*(4), 498–506. https://doi.org/10.1016/j.dss.2010.06.003

Delen, D., Topuz, K., & Eryarsoy, E. (2020). Development of a Bayesian Belief Network-based DSS for predicting and understanding freshmen student attrition. *European Journal of Operational Research, 281*(2), 575–587. https://doi.org/10.1016/j.ejor.2019.07.045

Delors, J. (1996). *Informe a la UNESCO de la Comisión Internacional sobre la Educación para el siglo XXI: La educación encierra un tesoro.* Madrid: Santillana, Ediciones UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000109590_spa

Dharmawan, T., Ginardi, H., & Munif, A. (2018). Dropout detection using non-academic data. *Proceedings - 2018 4th International Conference on Science and Technology, ICST 2018*, 1, 1–4. https://doi.org/10.1109/ICSTC.2018.8528619

Díaz de Cossío, R. (1998). Los desafíos de la educación superior mexicana. *Revista de la Educación Superior, 106*, 8. http://www.anuies.mx/index1024.html

Díaz, P., & Tejedor, A. (2017). El CADESU: Un nuevo instrumento para analizar la deserción estudiantil universitaria. *Reencuentro: Burocratización. https://reencuentro.xoc.uam.mx/index.php/reencuentro/article/view/925*

Dicovskiy, L. M., & Pedroza, M. E. (2018). Minería de datos, una innovación de los métodos cuantitativos de investigación, en la medición del rendimiento académico universitario. *Revista Científica De FAREM-Estelí, 24*, 143–152. https://doi.org/10.5377/farem.v0i24.5557

Digital55. (2020, junio 1). Qué es machine learning: Casos de éxito en empresas. *Digital55.*https://www.digital55.com/innovacion/que-es-machine-learning-casos-exito-empresas

Donoso, S., Donoso, G., & Arias, Ó. (2018). Iniciativas de retención de estudiantes de educación superior. *Calidad en la Educación, 33*, 15-61. https://doi.org/10.31619/caledu.n33.138

Eckert, K. B., & Suénaga, R. (2015). Análisis de deserción-permanencia de estudiantes universitarios utilizando técnica de clasificación en minería de datos. *Formación Universitaria, 8*(3), 3-12. https://doi.org/10.4067/S0718-50062015000300001

Esteban, M., Bernardo, A., Tuero, E., Cervero, A., & Casanova, J. (2017). Variables influyentes en progreso académico y permanencia en la universidad. *European Journal of Education and Psychology, 10*(2), 75–81. https://doi.org/10.1016/j.ejeps.2017.07.003

Fagella, D. (2020). Artificial intelligence in retail: 10 present and future use cases. *Emerj.* https://emerj.com/ai-sector-overviews/artificial-intelligence-retail

Friedman, J. (2001). *Greedy function approximation: A gradient boosting machine.* Stanford University. DOI:10.1214/aos/1013203451

García Ruiz de León, M. (2018). Análisis de sensibilidad mediante Random Forest. *Proyecto Fin de Carrera / Trabajo Fin de Grado, E.T.S.I. Industriales, UPM. https://oa.upm.es/53368/*

Garzón, A., & Gil, J. (2017). El papel de la procrastinación académica como factor de la deserción universitaria. *Revista Complutense de Educación, 28*(1), 307–324. https://doi.org/10.5209/rev_RCED.2017.v28.n1.49682

Girón, L., & González, D. (2005). Determinantes del rendimiento académico y la deserción estudiantil en el programa de economía de la Pontificia Universidad Javeriana de Cali. *Revista Kconomía, Gestión y Desarrollo, 3*, 173–201. https://bit.ly/3KJrGFv

Girones, J. M., & Casas, J. (2017). *Minería de datos: Modelos y algoritmos.* Editorial UOC.

Gracia, M. (2015). Deserción universitaria en México. *Milenio.* https://www.milenio.com/opinion/maximiliano-gracia-hernandez/la-economia-del-tunel/desercion-universitaria-en-mexico

González, L. E. (2005). Estudio sobre la repitencia y deserción en la educación superior chilena. *Digital Observatory for Higher Education in Latin America and The Caribbean, IESALC-UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000140087*

Gutiérrez, J. A., Garzón, J., & Segura, A. M. (2021). Factores asociados al rendimiento académico en estudiantes universitarios. *Formación Universitaria, 14*(1), 13–24. https://doi.org/10.4067/S0718-50062021000100013

Hämäläine, W., & Vinni, M. (2011). Clasificadores para la minería de datos educativos. En *Chapman & Hall/CRC* (pp. 57–74). https://repositorio.unican.es/xmlui/bitstream/handle/10902/8551/Tesis%20DGS.pdf

Han, J., & Kamber, M. (2006). *Data mining: concepts and techniques.* Morgan Kaufmann. https://doi.org/10.1016/C2009-0-61819-5

Heppen, J. B., & Bowles, S. (2008). Developing early warning systems to identify potential high school dropouts. *National High School Center, American Institutes for Research*, 1–13. https://eric.ed.gov/?id=ED521558

Hernádez, R., Fernandez, C., & Baptista, M. D. P. (2014). *Metodología de la investigación.* McGraw-Hill / Interamericana de Editores, S.A. de C.V.

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (2010). *Metodología de la investigación* (5.ª ed.). McGraw Hill Interamericana.

Heublein, U. (2013). Student drop-out from German higher education institutions. *European Journal of Education, 49*(4), 497–513. https://doi.org/10.1111/ejed.12097

Huerta, P., Velasco, M., & Jiménez, M. (2016). Causas de la deserción escolar en las telesecundarias de la zona 55. *Revista Huella de la Palabra, 8*.

Ino, R. P. (2008). ¿Qué es inteligencia artificial? *Revista de Información, Tecnología y Sociedad, 4.*

Ishitani, T., & Desjardins, S. (2002). A longitudinal investigation of dropout from college in the United States. *Journal of College Student Retention, 4*(2), 173–201.

John, G. H., & Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. *Eleventh Conference on Uncertainty in Artificial Intelligence*, 338–345.

Jović, A., Brkić, K., & Bogunović, N. (2015). A review of feature selection methods with applications. *Proceedings of the 38th International Convention on Information Communication Technology Electronics and Microelectronics*, 1200–1205.

Karegowda, A. G., Manjunath, A. S., & Jayaram, M. A. (2010). Feature subset selection problem using wrapper approach in supervised learning. *International Journal of Computer Applications, 1*(1), 13–17. https://doi.org/10.5120/100-197

Kemper, L., Vorhoff, G., & Wigger, B. U. (2020). Predicting student dropout: A machine learning approach. *European Journal of Higher Education, 10*(1), 28-47. https://doi.org/10.1080/21568235.2020.1718520

Kerlinger, F. N. (2002). *Investigación del comportamiento.* McGraw-Hill.

Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2004). Predicting students' performance in distance learning using machine learning techniques. *Applied Artificial Intelligence, 18*(5), 411–426. https://doi.org/10.1080/08839510490442058

Kubat, M. (2017). *An introduction to machine learning*. Springer.

Kumari, M., & Sunila, G. (2011). Comparative study of data mining classification methods in cardiovascular disease prediction. *International Journal of Computer Applications, 2*(1), 1–6.

Kuz, A., & Morales, R. (2023). Ciencia de datos educativos y aprendizaje automático: Un caso de estudio sobre la deserción estudiantil universitaria en México. *Education in the Knowledge Society (EKS)*, *24*, e30080.

Larsen, M., Sommersel, H., & Larsen, M. (2013). *Evidence on dropout phenomena at universities*. Danish Clearinghouse for Educational Research. http://edu.au.dk/fileadmin/edu/Udgivelser/Clearinghouse/Review/Evidence_on_dropout_from_universities_brief_version.pdf

Latif, A., Ai, C., & Aa, H. (2015). Economic effects of student dropouts: A comparative study. *Journal of Global Economics, 3*(2), 2–5. https://doi.org/10.4172/2375-4389.1000137

Lee, S., & Chung, J. Y. (2019). The machine learning-based dropout early warning system for improving the performance of dropout prediction. *Applied Sciences, 9*(15), 3093. https://doi.org/10.3390/app9153093

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2018). Feature selection: A data perspective. *ACM Computing Surveys, 50*(6), 1–45. https://doi.org/10.1145/3136625

Liem, J., Dillon, C., & Gore, S. (2001). Mental health consequences associated with dropping out of high school. *Annual Conference of the American Psychological Association*, 109. https://eric.ed.gov/?id=ED457502

Liz, C. (2011, March). Deserción estudiantil; Problema internacional, nacional, estatal y regional. Deserción universitaria. *Estres en Deserción Universitaria*. http://estresendesercionuniversitaria.blogspot.com/2011/03/antecedentes.html

López de Ullibarri, G. I., & Píta Fernández, S. (1998). Curvas ROC. *Cad Aten Primaria, 5*(4), 229–235.

Lugo, A. (2020). ¿Qué es el machine learning? *INVID*. https://invidgroup.com/es/machine-learning-metodos

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

Markov, Z., & Larose, D. T. (2007). *Data mining the Web: Uncovering patterns in Web content, structure, and usage*. John Wiley & Sons.

Márquez-Vera, C., Romero Morales, C., & Ventura Soto, S. (2013). Predicting school failure and dropout by using data mining techniques. *Revista Iberoamericana de Tecnologías del Aprendizaje, 8*(1), 7–14. https://doi.org/10.1109/RITA.2013.2244692

Martínez, A., Hernández, L. I., Carillo, D., Romualdo, Z., & Hernández, C. P. (2013). Factores asociados a la reprobación estudiantil en la Universidad de la Sierra Sur, Oaxaca. *Temas de Ciencia y Tecnología, 17*(51), 25–33. https://www.utm.mx/edi_anteriores/temas51/T51_1Ensayo3-FactAsocReprobacion.pdf

Microsoft Azure. (2022). *How to select algorithms for Azure machine learning*. Microsoft Azure. https://docs.microsoft.com/en-us/azure/machine-learning/how-to-select-algorithms

Mishra, A. (2018). Metrics to evaluate your machine learning algorithm. *Towards Data Science*. https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.

Ortega, F. (2012). *Los desertores del futuro* (Vol. 5). Centro de Estudios Avanzados: Universidad Nacional de Córdoba.

Oviedo, E., Oviedo, A. I., & Vélez, G. L. (2015). Minería de datos: Aportes y tendencias en el servicio de salud de ciudades inteligentes. *Revista Politécnica, 11*(20), 111–120.

Ozga, J., & Sukhnandan, L. (1998). Undergraduate non-completion: Developing an explanatory model. *Higher Education Quarterly, 52*(3), 316–333.

Páramo, G. J., & Correa, C. A. (1999). *Deserción estudiantil universitaria*. Universidad EAFIT. https://repositorio.unal.edu.co/xmlui/handle/unal/1075

Pek, R. Z., Özyer, S. T., Elhage, T., Özyer, T., & Alhajj, R. (2021). The role of machine learning in identifying students at-risk and minimizing failure. *IEEE Access, 11*, 1224–1243. https://doi.org/10.1109/ACCESS.2022.3232984

Peña, A. (2014). Review: Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications, 41*(4), 1432–1462. https://doi.org/10.1016/j.eswa.2013.08.042

Pérez-Planells, L., Delegido, J., Rivera-Caicedo, J., & Verrelst, J. (2016). Análisis de métodos de validación cruzada para la obtención robusta de parámetros biofísicos. *Universitat Politècnica de València*. http://hdl.handle.net/10251/80558

Pérez, A., Grandón, E. E., Caniupán, M., & Vargas, G. (2019). Comparative analysis of prediction techniques to determine student dropout: Logistic regression vs decision trees. *Proceedings of the International Conference of the Chilean Computer Science Society*. https://doi.org/10.1109/SCCC.2018.8705262

Peterson, L. E., & Coleman, M. A. (2008). Machine learning-based receiver operating characteristic (ROC) curves for crisp and fuzzy classification of DNA microarrays in cancer research. *International Journal of Approximate Reasoning, 47*(1), 17–36.

Piedra, J. A. M., & Manqueros, J. M. C. (2021). El muestreo y su relación con el diseño metodológico de la investigación. En *Manual de temas nodales de la investigación cuantitativa. Un abordaje didáctico* (p. 81).

Pineda, B., De Alvarado, E. L., & De Canales, F. (1994). *Metodología de la investigación: Manual para el desarrollo de personal de salud* (2da ed.). Organización Panamericana de la Salud.

Pradeep, A., Das, S., & Kizhekkethottam, J. J. (2015, febrero 25-27). Students dropout factor prediction using EDM techniques. En *Proceedings of the IEEE International Conference*

on Soft-Computing and Network Security (ICSNS 2015), Coimbatore, India. https://doi.org/10.1109/ICSNS.2015.7292406

Probst, P., Boulesteix, A.-L., & Bischl, B. (2018). Tunability: Importance of Hyperparameters of Machine Learning Algorithms. *Journal of Machine Learning Research, 20*, 53:1–53:32. https://doi.org/10.1007/s11276-017-1542-1

Python Software Foundation. (2021). *History and license*. Python 3 documentation. https://docs.python.org/3/license.html

Pulso. (2018, enero 10). Deserción, uno de los problemas vigentes en el sector educativo en México. *Pulso*. https://pulsoslp.com.mx/slp/desercion-uno-de-los-problemas-vigentes-en-el-sistema-educativo-de-mexico-nava/767958

Ramasubramanian, K. S. (2019). *Machine learning using R*. New Delhi, India: Karthik Ramasubramanian and Abhishek Singh.

Rico Páez, A., & Gaytán Ramírez, N. D. (2022). Modelos predictivos del rendimiento académico a partir de características de estudiantes de ingeniería. *IE Revista De Investigación Educativa De La REDIECH, 13*, e1426. https://doi.org/10.33010/ie_rie_rediech.v13i0.1426

Riquelme, J., Ruiz, R., & Gilbert, K. (2006). Minería de datos: Conceptos y tendencias. *Revista Iberoamericana de Inteligencia Artificial, 10*(29), 11-18.

Rochin Berumen, F. L. (2021). Deserción escolar en la educación superior en México: Revisión de literatura. *RIDE Revista Iberoamericana para la Investigación y el Desarrollo Educativo, 11*(22). https://doi.org/10.23913/ride.v11i22.821

Rodríguez Gómez, D. (2023). *Análisis de datos y machine learning* [Trabajo de fin de grado, Universidad de Salamanca]. Tutor: J. L. Vicente Villardón. https://docs.python.org/3/license.html

Rodríguez Lagunas, J., & Hernández Vázquez, J. (2008). La deserción escolar universitaria en México: La experiencia de la Universidad Autónoma Metropolitana campus Iztapalapa. *Actualidades Investigativas en Educación, 8*(3), 1-19.

Rodríguez, R. I. G., & Mazariego, C. R. R. (2020). Factores de deserción en los estudiantes de la licenciatura en Derecho, generación 2016-2020, en la Universidad Juárez Autónoma de Tabasco, División Académica Multidisciplinaria de los Ríos. *European Scientific Journal, ESJ, 16*(13), 54.

Romero, C., & Ventura, S. (2007). Minería de datos educativos: Un estudio de 1995 a 2005. *Sistemas Expertos con Aplicaciones, 1*, 135-146.

Romero, C., & Ventura, S. (2010). Minería de datos educativos: Una revisión del estado del arte. *IEEE Transactions on Systems, Man, and Cybernetics: Part C, 40*, 601-618.

Romero, C., & Ventura, S. (2013). Minería de datos en educación. *WIREs Data Mining and Knowledge Discovery, 3*, 12-27.

Romero, C., Espejo, P., Zafra, A., Romero, J., & Ventura, S. (2013). Minería de uso de la web para la predicción de notas de alumnos que utilizan cursos en Moodle. *Computer Applications in Engineering Education, 21*(1), 135-146. https://doi.org/10.1002/cae.20547

Ruiz, L. (2013). Aumenta deserción escolar en la UJAT. *Rumbo Nuevo*, p. 1.

Salal, Y. K., Abdullaev, S. M., & Kumar, M. (2019). Educational data mining: Student performance prediction in academic. *International Journal of Engineering and Advanced Technology, 8*(4C), 54-59. https://www.researchgate.net/publication/332369964_Educational_Data_Mining_Student_Performance_Prediction_in_Academic

Sánchez, G., Navarro, W., & García, A. (2009). Factores de deserción estudiantil en la Universidad Surcolombiana. *Paideia Surcolombiana, 1*(14), 97-103. https://doi.org/10.25054/01240307.1083

Segura-Morales, M., & Loza-Aguirre, E. (2018). Using decision trees for predicting academic performance based on socio-economic factors. In *Proceedings - 2017 International Conference on Computational Science and Computational Intelligence, CSCI 2017* (pp. 1132–1136). https://doi.org/10.1109/CSCI.2017.197

Segura, M., Mello, J., & Hernández, A. (2022). Machine learning prediction of university student dropout: Does preference play a key role? *Mathematics, 10*(18), 3359. https://doi.org/10.3390/math10183359

Selltiz, C., Wrightsman, L. S., & Cook, S. W. (1980). *Métodos de investigación en las relaciones sociales*. Ediciones RIALP S.A.

Shaikh, T. (2014). A prototype of Parkinson's and primary tumor diseases prediction using data mining techniques. *International Journal of Engineering Science and Invention, 3*(4).

Sierra Bravo, R. (2003). *Técnicas de investigación social: Teoría y ejercicios*. Thomson Editores.

Solís, M., Moreira, T., González, R., Fernández, T., & Hernández, M. (2018). Perspectivas para predecir la deserción escolar en estudiantes universitarios con machine learning. In *2018 IEEE International Work Conference on Bioinspired Intelligence, IWOBI 2018 - Proceedings* (pp. 1-8). https://doi.org/10.1109/IWOBI.2018.8464191

Székely, M. (2015). *Estudio sobre los principales resultados y recomendaciones de la investigación y evaluación educativa en el eje de prevención y atención a la deserción escolar en educación media superior*. Instituto Nacional para la Evaluación de la Educación (INEE).

Talamás-Carvajal, J. A., & Ceballos, H. G. (2023). A stacking ensemble machine learning method for early identification of students at risk of dropout. *Educational Information Technology, 28*, 12169–12189. https://doi.org/10.1007/s10639-023-11682-z

Talavera, L. (2005). An evaluation of filter and wrapper methods for feature selection in categorical clustering. In *Proceedings of the International Symposium on Intelligent Data Analysis* (pp. 440-451). Madrid, Spain.

Tavico, A. (2021). Factores que influyen en la deserción de la carrera de administración de empresas del centro universitario de Quiché - CUSACQ. *Revista Científica Internacional, 4*(1), 39-46. https://doi.org/10.46734/revcientifica.v4i1.45

Timarán-Pereira, I., Hernández-Arteaga, S. R., Caicedo-Zambrano, S. J., Hidalgo-Troya, A., & Alvarado-Pérez, J. C. (2016). El proceso de descubrimiento de conocimiento en bases de datos. In *Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional* (pp. 63-86). Ediciones Universidad Cooperativa de Colombia.

Timarán, R., Calderón, A., & Jiménez, J. (2013). Descubrimiento de perfiles de deserción estudiantil con técnicas de minería de datos. *Revista Vínculos, 10*(1), 373-383. https://doi.org/10.14483/2322939X.4687

Tinto, V. (1982). Limits of theory and practice in student attrition. *The Journal of Higher Education, 53*(6), 687-700. https://doi.org/10.1080/00221546.1982.11778398

Tinto, V. (1989). Definir la deserción: Una cuestión de perspectiva. *Revista de la Educación Superior, 71*, 33-35.

Tinto, V. (1989). Una reconsideración de las teorías de la deserción estudiantil. *Trayectoria escolar en la educación superior*, 47-84.

Tinto, V. (2006). Research and practice of student retention: What next? *Journal of College Student Retention: Research, Theory & Practice, 8*(1), 1-19. https://doi.org/10.2190/3W6W-7G1W-0M8F-861K

Torres, P. C., & Cobo, J. K. (2017). Tecnología educativa y su papel en el logro de los fines de la educación. *Educere, 21*(68), 31-40. https://dialnet.unirioja.es/servlet/articulo?codigo=6560961

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics, 33*(1), 1-67. https://doi.org/10.1214/aoms/1177704711

Valero Cajahuanca, J. E., Navarro Raymundo, A. F., Larios Franco, A. C., & Julca Flores, J. D. (2022). Deserción universitaria: Evaluación de diferentes algoritmos de machine learning para su predicción. *Revista de Ciencias Sociales, 28*(3), 362-375. https://doi.org/10.31876/rcs.v28i3.38480

Vialardi, C., Chue, J., Peche, J. P., Alvarado, G., Vinatea, B., Estrella, J., & Ortigosa, J. (2011). Un enfoque de minería de datos para guiar a los estudiantes a través del proceso de inscripción basado en el rendimiento académico. *User Modeling and User-Adapted Interaction, 21*(1-2), 217-248.

Vries, W. D., León, P., Romero, J. F., & Hernández, I. (2011). ¿Desertores o decepcionados? Distintas causas para abandonar los estudios universitarios. *Revista de la Educación Superior, 40*(160), 29-49. https://doi.org/10.22201/iisue.01875357e.2011.160.153

Weiss, E. (2015). El abandono escolar en la educación media superior: Dimensiones, causas y políticas para abatirlo. En R. Ramírez (Ed.), *Desafíos de la educación media superior* (pp. 81-160). Instituto Belisario Domínguez.

Wolff, A., Zdrahal, Z., Herrmannova, D., & Knoth, P. (2014). Predicción del rendimiento de los estudiantes a partir de fuentes de datos. *User Modeling and User-Adapted Interaction, 21*(1-2), 217-248.

Xiaojin, Z. (2008). *Semi-Supervised Learning Literature Survey*. University of Wisconsin, Madison.

Yoo, J., & Kim, J. (2014). ¿Puede la participación en debates en línea predecir el rendimiento de un proyecto de grupo? Investigating the roles of linguistic features and participation patterns. *Revista Internacional de Inteligencia Artificial en la Educación, 24*(1), 8-32.

Yoti, E., & Walia, E. A. S. (2017). A review on recommendation system and web usage data mining using k-nearest neighbor (KNN) method. *International Research Journal of Engineering and Technology (IRJET), 4*(4), 2931-2934.

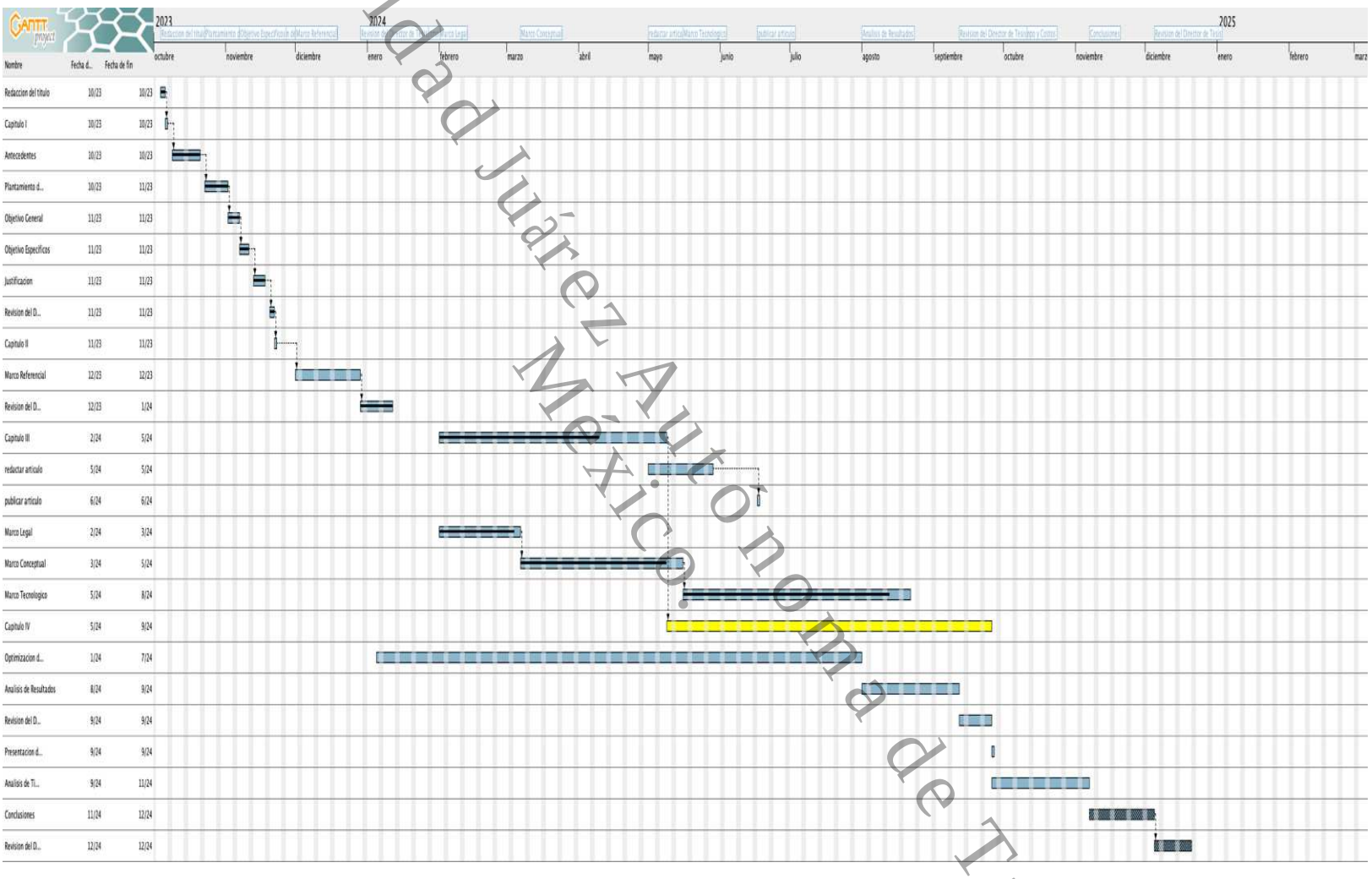
Youssef, M., Mohammed, S., Hamada, E. K., et al. (2019). A predictive approach based on efficient feature selection and learning algorithms' competition: Case of learners' dropout in MOOCs. *Educ Inf Technol, 24*, 3591–3618. https://doi.org/10.1007/s10639-019-09934-y

Zambrano, J. (2021). Efectos de la pandemia del COVID-19 en la deserción de estudiantes de Ciencias de la Comunicación, Universidad Científica del Perú, San Juan 2020 – I [Tesis de maestría, Universidad Científica del Perú]. Repositorio Institucional Digital. https://bit.ly/3dE4LPp

Zapata-Medina, D., Espinosa-Bedoya, A., & Jiménez-Builes, J. A. (2024). Improving the automatic detection of dropout risk in middle and high school students: A comparative study of feature selection technique. *Mathematics, 12*(1776). https://doi.org/10.3390/math12121776

Zapeta Hernández, A., Galindo Rosales, G. A., Juan Santiago, H. J., & Martínez Lee, M. (2022). Métricas de rendimiento para evaluar el aprendizaje automático en la clasificación de imágenes petroleras utilizando redes neuronales convolucionales. *Ciencia Latina Revista Científica Multidisciplinar, 6*(5), 4624-4637. https://doi.org/10.37811/cl_rcm.v6i5.3420

6.1. Anexo 1 Cronograma de actividades



6.2. Anexo 2 Alojamiento de la Tesis en el Repositorio Institucional

Alojamiento de la Tesis en el Repositorio Institucional	
Título de Tesis:	Identificación de la deserción escolar usando técnicas de selección de características en modelos de aprendizaje automático.
Autor(a) o autores(ras) de la Tesis:	Daniel Domínguez Gómez
ORCID:	https://orcid.org/0009-0007-4695-1978
Resumen de la Tesis:	Este estudio aborda el reto de identificar a los estudiantes que abandonan los estudios centrándose en la selección de características o variables relevantes. Aplicamos cinco técnicas de selección de características: ANOVA, información mutua, selección secuencial hacia delante, eliminación recursiva de características y el operador de selección y reducción mínima absoluta, y destacamos las características sociodemográficas clave. Los resultados mostraron que el modelo KNN de información mutua obtuvo la precisión y la puntuación F1 más altas (99,05 %), mientras que SFS proporcionó resultados óptimos para los árboles de decisión y las máquinas de vectores de apoyo. Estos resultados ponen de manifiesto el papel fundamental de las técnicas de selección de características a la hora de identificar las variables que afectan significativamente a la eficacia de los modelos predictivos.
Palabras claves de la Tesis:	abandono escolar, predicción, educación, técnicas de selección de características, aprendizaje automático, algoritmos.

Referencias citadas:	Se muestra a partir de la página 82.