FACULTAD DE POSTGRADO
TRABAJO FINAL DE GRADUACIÓN

ANÁLISIS PREDICTIVO DEL RIESGO DE DESERCIÓN EN ESTUDIANTES DE PRIMER INGRESO DE LA UNIVERSIDAD TECNOLÓGICA CENTROAMÉRICANA DE LOS AÑOS 2023 AL 2025

SUSTENTADO POR:

DAVID ELÍAS FLORES MALDONADO
WILMER ALEXANDER VÁSQUEZ RODRÍGUEZ

PREVIA INVESTIDURA AL TÍTULO DE

MÁSTER EN
GESTIÓN DE TECNOLOGÍAS DE LA INFORMACIÓN

TEGUCIGALPA, FRANCISCO MORAZÁN, HONDURAS, C.A.

ENERO, 2026

UNIVERSIDAD TECNOLÓGICA CENTROAMERICANA

UNITEC

FACULTAD DE POSTGRADO

AUTORIDADES UNIVERSITARIAS

RECTORA

ROSALPINA RODRÍGUEZ

VICERRECTOR ACADÉMICO NACIONAL

JAVIER ABRAHAM SALGADO LEZAMA

SECRETARIO GENERAL

ROGER MARTÍNEZ MIRALDA

DECANA FACULTAD DE POSTGRADO

ANA DEL CARMEN RETTALLY VARGAS

ANÁLISIS PREDICTIVO DEL RIESGO DE DESERCIÓN EN ESTUDIANTES DE PRIMER INGRESO DE LA UNIVERSIDAD TECNOLÓGICA CENTROAMÉRICANA DE LOS AÑOS 2023 AL 2025

TRABAJO PRESENTADO EN CUMPLIMIENTO DE LOS REQUISITOS EXIGIDOS PARA OPTAR AL TÍTULO DE

MÁSTER EN

GESTIÓN DE TECNOLOGÍAS DE LA INFORMACIÓN

ASESOR

JESÚS RICARDO RODRÍGUEZ RIVERA

MIEMBROS DE LA TERNA:

ELVIN BOBADILLA
JOSÉ LUIS MARTÍNEZ
ANTHONY BARAHONA

FACULTAD DE POSTGRADO

ANÁLISIS PREDICTIVO DEL RIESGO DE DESERCIÓN EN ESTUDIANTES DE PRIMER INGRESO DE LA UNIVERSIDAD TECNOLÓGICA CENTROAMÉRICANA DE LOS AÑOS 2023 AL 2025

David Elías Flores Maldonado
Wilmer Alexander Vásquez Rodríguez

Resumen

El presente trabajo tiene la finalidad efectuar un análisis predictivo del riesgo de deserción en estudiantes de primer ingreso de la Universidad Tecnológica Centroamericana de los años 2023 al 2025, mediante del entrenamiento de los modelos *Gradient Boosted Trees Random Forest*, *Decision Tree, K-Nearest Neighbors* y Regresión Logística, usando *Stratified K-Fold Cross Validation* para garantizar una evaluación justa, midiendo cada modelo con métricas comunes (exactitud, *recal*l, precisión*, F1-score y Cohen's kappa*) y matrices de confusión creadas en KNIME. Todo ello con el propósito de identificar tempranamente a los estudiantes en riesgo académico y actuar oportunamente. Ya que la aplicación de herramientas de análisis predictivo permite tomar decisiones más acertadas y oportunas para brindar un eficiente acompañamiento a los estudiantes. Este es un elemento clave que genera una ventaja competitiva al contar con una gestión más eficiente de las bases de datos y registros de los estudiantes. A través de los resultados se demostró que el modelo *Gradient Boosted Trees* tiene un desempeño robusto en la predicción del riesgo de deserción estudiantil, evidenciando una alta capacidad discriminativa, clasificando correctamente a 13,711 estudiantes (88.48% del total).

Palabras claves: (Acompañamiento estudiantil, Anticipación, Aprendizaje automático, Minería de datos, Modelos de predicción)

GRADUATE SCHOOL

PREDICTIVE ANALYSIS OF DROPOUT RISK AMONG FIRST-YEAR STUDENTS AT THE CENTRAL AMERICAN TECHNOLOGICAL UNIVERSITY FROM 2023 TO 2025

David Elías Flores Maldonado
Wilmer Alexander Vásquez Rodríguez

Abstract

The purpose of this study is to conduct a predictive analysis of the risk of dropout among first-year students at the Central American Technological University during the period 2023–2025. This is achieved through the training of Gradient Boosted Trees, Random Forest, Decision Tree, K-Nearest Neighbors, and Logistic Regression models, using Stratified K-Fold Cross Validation to ensure a fair and robust evaluation. Each model is assessed using common performance metrics (accuracy, recall, precision, F1-score, and Cohen's kappa), as well as confusion matrices generated in KNIME. The overarching objective is to enable the early identification of students at academic risk and to support timely intervention. The application of predictive analytics tools facilitates more accurate and timely decision-making, allowing institutions to provide effective academic support to students. This approach represents a key element in achieving a competitive advantage by enabling more efficient management of student databases and academic records.

The results demonstrate that the Gradient Boosted Trees model exhibits robust performance in predicting student dropout risk, showing high discriminative capability and correctly classifying 13,711 students (88.48% of the total population).

Keywords: (Data mining, Early detection, Machine learning, Predictive models, Student support)

DEDICATORIA

A Dios sobre todas las cosas, mis padres por su continuo amor y apoyo incondicional, a mi tía Nora, a mi tío William y mi prima Norali que, aunque se encuentren lejos me han brindado su amor y cariño, a mis abuelos que en paz descansen, todos me han brindado la motivación para seguir adelante.

David Elías Flores Maldonado

En primer lugar, lo Dios, por la sabiduría y fortaleza para culminar mi trabajo de tesis, uno de mis proyectos de vida. También a mis padres por todo su apoyo, amor, comprensión y ánimo para poder seguir adelante en este proceso difícil, pero al mismo tiempo muy satisfactorio para mi vida profesional y personal.

En general a toda mi familia y amigos, por todo su cariño y ánimo para seguir adelante y poder cumplir esta gran meta.

Wilmer Alexander Vásquez Rodríguez

AGRADECIMIENTO

A la Universidad Tecnológica Centroamericana, por brindarnos la oportunidad de contar con la información necesaria para realizar este proceso en su prestigiosa institución.

A todos los docentes que formaron parte de todo el proceso de la Maestría en Gestión de Tecnologías de la Información, haciendo mención especial a nuestro asesor, Máster Jesús Ricardo Rodríguez Rivera, por su continua dedicación y apoyo en la elaboración de esta tesis, con aportes relevantes y precisos desde el inicio hasta el final de nuestro informe, para permitirnos subir un peldaño más en cada una de sus intervenciones.

Muchas gracias

ÍNDICE DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

CAPÍTULO I. PLANTEAMIENTO DE LA INVESTIGACIÓN

1.1 INTRODUCCIÓN

En el ámbito de la educación superior, el bajo rendimiento académico es el resultado de múltiples factores y causas, en este sentido los modelos predictivos permiten anticipar el riesgo académico con base en datos históricos y patrones comportamientos. Según el Instituto Nacional de Estadística (INE, 2023) en un país como Honduras, donde los indicadores de pobreza se sitúan en 64.1%, sumado a ello según un estudio realizado por la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO, 2021), el porcentaje de estudiantes que ingresa a estudiar a las universidades es de tan solo un 17%, por lo que, no es raro que la calidad del sistema educativo sea deficiente y la competitividad de las instituciones sea baja.

En el pasado este tipo de análisis del desempeño académico se ha realizado con métodos tradicionales, los cuales, si cumplen con su propósito, pero no resultan suficientes para evitar de manera temprana los riesgos que enfrentan los estudiantes, el uso de la minería de datos nos sirve como una vía para manejar o transformar grandes volúmenes de información con la cual podremos identificar patrones ocultos y factores de riesgo asociados al bajo rendimiento.

Mediante los algoritmos de clasificación, predicción y agrupación podemos prevenir riesgos, pero también intervenir, optimizar, analizar y fortalecer los programas de ayuda académica para beneficiar a los estudiantes.

Con este estudio se busca analizar una de las universidades privadas más populares de Honduras, con el propósito de conseguir una muestra lo suficientemente efectiva para poder utilizarla en el modelo predictivo que contribuya a la mejora del índice académico de las instituciones educativas.

Este estudio se estructura en 6 capítulos, el primer capítulo expone el planteamiento, abordando los escenarios donde ocurre el problema de deserción estudiantil y nos muestra la relevancia del tema, el segundo capítulo expone el marco teórico abordando los conceptos claves sobre rendimiento académico, minería de datos y modelos de predicción aplicados a la educación superior. El tercer capítulo presenta la metodología en al cual se detalla el enfoque, el tipo de estudio, la población, la muestra y los métodos de recolección de datos, además del análisis de

datos. En el cuarto capítulo se muestran los resultados y análisis, así como las evaluaciones de los modelos predictivos propuestos. El quinto capítulo exhibe las conclusiones y recomendaciones orientas a fortalecer las estrategias de apoyo académico. El sexto capítulo presenta la aplicabilidad que consiste en mostrar cómo y dónde se puede usar la práctica de los resultados obtenidos por la minería de datos.

1.2 ANTECEDENTES DEL PROBLEMA

En la actualidad la educación superior ha tenido bastante auge, sobre todo en las instituciones privadas, gracias a la flexibilidad de horarios y apertura de diversas modalidades que permiten a los estudiantes gestionar su aprendizaje de una forma muy diferente a la tradicional. Sin embargo, esto no significa que se esté exento de que los estudiantes en algún momento se encuentren en riesgo académico. (Carpio et al., 2018) como se citó en Guzmán-Torres et al., (2022) define el riesgo académico como "una condición de propensión cuya actualización puede adoptar la forma de rezago escolar, bajo nivel de aprovechamiento académico, bajo rendimiento escolar, o en el peor de los casos de fracaso escolar".

Para cada una de estas condiciones se puede aplicar la analítica del aprendizaje que se define como, "un paradigma que consiste en la medición, recopilación, análisis e informe de datos sobre los estudiantes y sus contextos, con el fin de comprender y optimizar el aprendizaje y los entornos en los que se produce" (Contreras-Bravo et al., 2021). Gracias a la analítica del aprendizaje las universidades pueden realizar una mejor gestión de los datos que recopilan lo cual genera una toma decisiones basada en datos que beneficien tanto a la institución como a los mismos estudiantes; y de esta manera evitar los altos grados de deserción al proporcionar un acompañamiento oportuno a los estudiantes que se encuentran en riesgo académico.

Hoy en día el uso de las herramientas de minería de datos ha tomado mucho auge ya que las organizaciones pueden tomar decisiones informadas con base en datos y así lograr una mejor gestión de clientes, y una mejora continua en procesos. En este sentido, el ámbito educativo no se puede quedar atrás, ya que la implementación de este tipo de herramientas permite predecir eficientemente muchos patrones de los estudiantes.

(Urbina-Nájera et al., 2020) afirman que la aplicación de minería de datos en la educación "tiene como objetivo seguir la huella digital de los estudiantes y descubrir de manera oportuna un cambio en el comportamiento vinculado a aspectos académicos que puedan predecir, por ejemplo,

una inminente deserción o abandono escolar" (p. 2).

En este sentido, se puede potencializar la minería de datos en las universidades con el fin de lograr un mejor seguimiento de los estudiantes de primer ingreso, así como su rendimiento académico, el cual, en algún momento puede denotar el riesgo de deserción académica.

Por lo cual podemos definir la minería de datos educativos como "una disciplina relativamente nueva que busca desarrollar nuevas técnicas para examinar conjuntos de datos obtenidos de entorno educativo y aplicarlas para arrojar nueva luz sobre estudiantes y los entornos educativos". (Papadogiannis et al., 2024).

Gracias a la obtención de ese conjunto de datos sobre diversos aspectos educativos, se puede predecir los estudiantes que se encuentran en riesgo académico, no obstante, es necesario utilizar algoritmos predictivos como: árboles de decisión que "es una estructura similar a un árbol donde cada nodo interno representa una decisión basada en una característica, cada rama representa el resultado de dicha decisión y cada nodo que sería la hoja, representa una etiqueta de clase". (Papadogiannis et al., 2024).

Es importante destacar que en los algoritmos predictivos en la educación superior se debe tomar en cuenta aspectos éticos y de privacidad, Jaime (2023), cómo se citó en Jaramillo Flores, (2024), señala que el "uso de datos personales y académicos para la predicción del éxito estudiantil debe equilibrase cuidadosamente con la protección de la privacidad y la prevención de sesgos algorítmico" (p. 4). Es importante tener en consideración estos aspectos para lograr una predicción ética y evitar que existan contratiempos con la aplicación de los modelos predictivos.

La implementación de modelos predictivos en las universidades permite un análisis más detallado y oportuno de los elementos que determinen si un estudiante se encuentra o no en riesgo académico; y lograr un seguimiento efectivo con el propósito de evitar la deserción.

1.3 DEFINICIÓN DEL PROBLEMA

1.3.1 ENUNCIADO DEL PROBLEMA

En la actualidad se cuenta con diversas herramientas de análisis de datos que permiten determinar los factores que influyen en el riesgo académico y es uno de los pilares fundamentales para predecir el acompañamiento que cada estudiante requiere y proporcionar el apoyo oportuno y enfocado en las necesidades de cada uno. No obstante, en muchas universidades estos datos no

son utilizados de forma estratégica ya que existe la ausencia de análisis predictivo lo que restringe la posibilidad de que las autoridades académicas actúen de manera anticipada.

Es por ello, que es imperante analizar datos históricos de bases de datos académicos de los estudiantes, para poder identificar patrones que podrían prever el bajo desempeño o el riesgo de deserción de los estudiantes.

Este proyecto de investigación tiene el propósito de analizar el riesgo académico en estudiantes de primer ingreso mediante la aplicación de técnicas de minería de datos en la Universidad Tecnológica Centroamericana con el fin de contar con un análisis predictivo que permita un acompañamiento oportuno en los estudiantes y evitar la deserción.

1.3.2 FORMULACIÓN DEL PROBLEMA

En estudiantes de primer ingreso de la Universidad Tecnológica Centroamericana (P), ¿qué tan efectiva es la aplicación de técnicas de minería de datos y *learning analytics* (I) para la predicción y anticipación del riesgo de deserción (O), en comparación con el seguimiento académico tradicional (C)?

1.4 PREGUNTAS DE INVESTIGACIÓN

1. En estudiantes de primer ingreso (P), ¿en qué medida los factores más predictivos del riesgo de deserción (O), al ser analizados mediante modelos predictivos (árboles de decisión y regresión logística) (I), permiten una identificación más temprana y precisa del riesgo de deserción (C)?
2. En estudiantes de primer ingreso (P), ¿qué precisión predictiva y capacidad de anticipación demuestran los modelos predictivos (árboles de decisión y regresión logística) (I) para identificar el riesgo de deserción (O), en contraste con la capacidad de detección del seguimiento académico tradicional (C)?
3. En estudiantes de primer ingreso (P), ¿de qué manera la información generada por un modelo predictivo (árboles de decisión y regresión logística) (I) puede optimizar el diseño y la pertinencia de las estrategias de acompañamiento estudiantil para reducir el riesgo de deserción (O), en comparación con la información obtenida del seguimiento académico tradicional (C)?

1.5 OBJETIVOS DEL PROYECTO

1.5.1 OBJETIVO GENERAL

Evaluar (S) la efectividad de la aplicación de técnicas de minería de datos y *learning analytics para* la predicción y anticipación del riesgo de deserción en estudiantes de primer ingreso de la Universidad Tecnológica, (M) mediante indicadores de precisión y capacidad de anticipación, (A) utilizando datos institucionales disponibles, (R) en comparación con el seguimiento académico tradicional, y (T) dentro de unos períodos académicos definidos.

1.5.2 OBJETIVOS ESPECÍFICOS

1. Identificar (S) los factores más predictivos del riesgo de deserción en estudiantes de primer ingreso (M) mediante la aplicación de modelos predictivos como árboles de decisión y regresión logística, (A) utilizando datos históricos disponibles de la Universidad Tecnológica Centroamericana, (R) para una determinación más temprana y precisa de este, y (T) al finalizar el análisis de dichos datos históricos.
2. Determinar (S) la precisión predictiva y la capacidad de anticipación de los modelos de árboles de decisión y regresión logística para identificar el riesgo de deserción en estudiantes de primer ingreso, (M) cuantificando su desempeño frente al seguimiento académico tradicional, (A) utilizando datos del año académico de estudio, (R) para mejorar los mecanismos de detección temprana, y (T) dentro del marco del período académico analizado.
3. Analizar (S) cómo la información generada por modelos predictivos como árboles de decisión y regresión logística puede optimizar el diseño y la pertinencia de las estrategias de acompañamiento estudiantil, (M) mediante la formulación de recomendaciones concretas basadas en evidencia, (A) a partir del análisis comparativo con el seguimiento académico tradicional, (R) para reducir el riesgo de deserción en estudiantes de primer ingreso, y (T) posterior al procesamiento y evaluación de los resultados obtenidos.

1.6 JUSTIFICACIÓN

En esta investigación la justificación se construye sobre una base de pertinencia del problema identificado y lo solido de la metodología propuesta, asegurando que el estudio no solo sea relevante para las universidades, sino que también nos brinde un valor tangible en el mundo real. Siguiendo la premisa que una justificación solida se basa en datos y la realidad, y no solo en opiniones, por eso se abordaran las tres dimensiones del valor.

Un problema que se da en la mayoría de las instituciones educativas es el riesgo académico de los estudiantes ya que es un factor determinante en la calidad educativa y en la eficiencia de estas, la deserción, el bajo rendimiento y atraso de los estudios afecta el desarrollo profesional de los estudiantes y también genera un impacto económico y reputacional en las universidades. Según la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (Aveleyra, 2023), el índice de deserción estudiantil de las universidades en América Latina es del 40% lo que demuestra la magnitud de este problema, Actualmente se tiene acceso a grandes volúmenes de información, las instituciones tienen la oportunidad de utilizar sus propios datos académicos para visualizar e identificar de manera temprana a los estudiantes de alto riesgo y diseñar estrategias preventivas más efectivas.

El problema de la deserción estudiantil es una preocupación tangible y significativa para las instituciones de educación superior, ya que esto daña la imagen y reputación de las instituciones educativas. Implica no solo la pérdida de talento y la interrupción de trayectorias académicas individuales, sino también costos económicos y de recursos considerables para las universidades, la pérdida de ingresos continuos perjudica la calidad de educación. La gestión tradicional de este riesgo a menudo es reactiva, interviniendo cuando el problema ya está avanzado o es irreversible, En Honduras, solo el 9% de los jóvenes logra completar la educación superior (Arias, 2025), en este sentido, es primordial transformar la forma en que se gestiona la permanencia estudiantil, y ante la ausencia de un sistema de análisis predictivo que permita una actuación anticipada; este estudio busca aprovechar la minería de datos para identificar patrones de riesgo y lograr una gestión preventiva. Con el fin de proporcionar a la Universidad Tecnológica Centroamericana un sistema de alerta temprana que les permita pasar de una respuesta reactiva a una intervención preventiva y personalizada; que les permita reducir la deserción estudiantil y obtener una ventaja competitiva mediante una gestión eficiente de las bases de datos optimizando los recursos

disponibles.

Utilizando diversas técnicas de minería de datos avanzadas, podemos descubrir patrones ocultos en la información, analizar e identificar factores que determinan el rendimiento y que nos ayuden a predecir posibles escenarios de riesgo académico, mediante la utilización de técnicas, se facilita la integración de variables académicas, socioeconómicas y de comportamiento para elaborar modelos predictivos que faciliten la toma de decisiones basada en evidencias.

Este proyecto tiene como justificación la necesidad de contar con un sistema de análisis robusto que permita a las universidades no solo detectar casos de riesgo, sino también prevenir o anticiparse a estos riesgos, mediante la optimización de los programas de apoyo y reducción de deserción. La propuesta nos ayuda a fortalecer la gestión académica mediante un enfoque analítico que tiene aplicabilidad en universidades privadas de Honduras, puede adaptarse a otros contextos educativos, que ayuda a contribuir el avance de la analítica educativa de la región, para justificar lo anterior se demostrara la relevancia desde tres perspectivas, la académica, la social y la económica, estas dimensiones nos ayudaran a comprender el impacto integral de los riesgos académicos y la deserción estudiantil.

Dimensión Académica

Un factor determinante en la calidad educativa es la mejora del rendimiento de los estudiantes universitarios, ya que apoya la acreditación institucional y la reputación académica. Lo que tiene un gran impacto es el bajo rendimiento y la deserción ya que estas reducen la tasa de graduación y afectan directamente en los indicadores de calidad exigidos por entes reguladores. En este estudio se brindará a las universidades privadas de Honduras una herramienta basada en minería de datos que mediante varios modelos predictivos nos ayudaría a reaccionar ante estos riesgos con anticipación.

Dimensión Social

Una de las causas que genera la deserción y el bajo rendimiento no solo afecta de manera individual, interrumpiendo la trayectoria profesional, también impacta de manera negativa el entorno familiar y la capacidad del país para contar con profesionales cualificados, actualmente en Honduras el acceso a la educación es limitado y la pobreza es elevada, prevenir el abandono contribuye a mejorar la movilidad social y reducir desigualdades, El (PNUD, 2022) indica que el

acceso a la educación en Honduras es limitado y las condiciones de pobreza son elevadas, por lo que prevenir el abandono contribuye a mejorar la movilidad social y a reducir desigualdades estructurales.

Dimensión Económica

La universidad tiene un problema con la deserción de sus estudiantes ya que esto implica una pérdida de ingresos continua, costos asociados a procesos administrativos y potencial de daño reputacional. Al tener una falta de profesionales cualificados reduce la competitividad y la capacidad de innovación del país. Este estudio propone un sistema de predicción que optimizará los recursos, mejorará la asignación de programas de apoyo y generará una disminución en la deserción estudiantil.

Asimismo, según La Encuesta Permanente de Hogares del Instituto Nacional de Estadística (INE, 2023), las principales causas por la que no está estudiando este porcentaje de estudiantes son:

- Porque se encuentran laborando
- La escasez de recursos económicos
- La falta de motivación para seguir estudiando
- Realiza ayuda en los quehaceres del hogar.

El diagrama de Ishikawa fue elaborado para identificar los orígenes de la deserción estudiantil en estudiantes de primer ingreso, esto nos permitirá organizar varios factores determinantes en seis categorías las cuales son: personales, socioeconómicas, académicos, institucionales, familiares y tecnológicos, estos aspectos nos aportaran una perspectiva complementaria que en conjunto explica la complejidad de este caso.

Los factores académicos y personales influyen directamente en el rendimiento y la probabilidad de permanencia de los estudiantes de educación superior (Aveleyra, 2023), debido a contenidos que son difíciles comprender y la falta de hábitos de estudio.

Con los factores socioeconómicos y tecnológicos se empeora esta situación ya que muchos estudiantes deben trabajar o carecen de recursos tecnológicos, limitando su capacidad de participación durante las clases y debido a esto no cumplen con las exigencias académicas (INE,

2023a), la falta de acompañamiento institucional y la poca comunicación de los docentes incremente la probabilidad de abandono. (UNAH, 2025b)

En un entorno familiar y contextual también influye en las responsabilidades domésticas, baja valoración de educación en el hogar y un mercado laboral poco atractivo que fomenten la desmotivación (PNUD, 2022).

En conclusión, la deserción estudiantil es el resultado de la interacción de múltiples factores, según la Encuesta Permanente de Hogares del Instituto Nacional de Estadística (INE, 2023), una de las principales causas de no estudiar en Honduras son: trabajar, la falta de recursos económicos, desmotivación y responsabilidades, A la par de los factores desalentadores antes expuestos, se suma el del mercado laboral, donde el nivel de ocupados de los egresados de las universidades es menor al 50%, lo que causa desaliento en la población juvenil, misma que ve la producción en las redes sociales como YouTube y TikTok una fuente mayor de ingresos. Esto refleja diversos estudios realizados en varias partes del mundo, donde los jóvenes ya no sueñan con ser profesionales, sino productores de contenidos, según la (UNAH, 2023), un sistema predictivo basado en minería de datos puede ayudar a identificar estos riesgos de manera temprana y optimizar los recursos universitarios para reducir significativamente la deserción (PNUD, 2022).

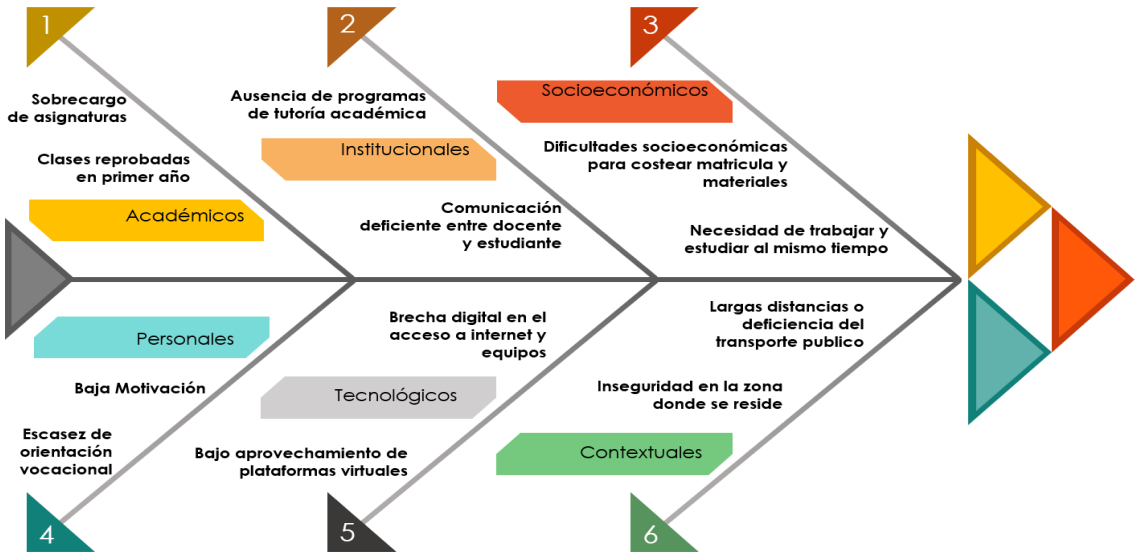


Ilustración 1 - Análisis de la alta deserción en estudiantes de primer ingreso con materias reprobadas

CAPÍTULO II. MARCO TEÓRICO

El objetivo de este capítulo es dar a conocer los fundamentos teóricos y conceptuales que sirven de respaldo a la investigación sobre del análisis predictivo para el riesgo de deserción en estudiantes de primer ingreso de la Universidad Tecnológica Centroamericana. Con el propósito de comprender el problema de investigación, es primordial realizar un análisis contextual que facilite explorar las interrelaciones de los elementos que influyen en la deserción estudiantil.

2.1 ANÁLISIS DEL MACROENTORNO

Se ha decidido un análisis PESTEL con México, Colombia y España para contrastar con la situación en Honduras. A pesar de que estos países tienen diferentes niveles de desarrollo, poseen rasgos históricos, culturales o socio-tecnológicos comunes que los hacen modelos de referencia provechosos para este análisis, debido al conjunto de lecciones únicas y complementarias que pueden proporcionar.

México:

Político (P): Por la aplicación de reformas significativas, entre 2018 y 2024, México sufrió cambios en sus políticas educativas, lo que modificó el marco regulador de la educación superior. “El Programa Sectorial de Educación 2020-2024 ha sentado las bases esenciales para financiar y desarrollar la educación superior, al instaurar un nuevo enfoque que da prioridad a la obligatoriedad y a la gratuidad progresiva de este nivel educativo” (Casillas Alvarado et al., 2021a). Esto ha hecho posible que toda la población tenga acceso a la educación superior, el cual está determinado por los elementos políticos de la nación.

Los cimientos normativos de esta transformación se han basado en la reforma constitucional del artículo 3° que se llevó a cabo en 2019 y la promulgación de la Ley General de Educación Superior (LGES) en abril del año 2021. La LGES determina que “el Estado es el responsable de la obligatoriedad de la educación superior y la define como un derecho que contribuye al desarrollo y bienestar integral del ser humano” (Lissen & Bautista, 2021).

Esta legislación indica una transición hacia políticas de gobierno que han avanzado hacia un enfoque más inclusivo, priorizando el desarrollo integral del ser humano por encima del entrenamiento técnico.

El marco constitucional reformado sostiene que la educación brindada por el Estado tiene que ser "inclusiva, pública, universal, gratuita y laica, lo cual significa un cambio de paradigma en cómo se concibe a la educación superior como un derecho esencial" (Reforma 2019 a los artículos 3°, 31 y 73 de la Constitución Política de los Estados Unidos Mexicanos, 2019). Las autoridades locales y federales han diseñado políticas específicas para promover la inclusión, el acceso universal y la equidad, a fin de consolidar un sistema educativo fundado en el respeto total de la dignidad humana, con un enfoque de igualdad y derechos humanos.

Las políticas de las universidades en relación con la digitalización se mantienen divididas y enfocadas en una adopción parcial de la tecnología. El análisis detallado de instituciones, como la Universidad de Guadalajara, muestra que "las políticas universitarias están divididas y se centran únicamente en integrar tecnología a una parte de sus procesos institucionales. Esto pone de manifiesto la necesidad de contar con marcos referenciales gubernamentales más precisos y completos" (Ramírez-Díaz, 2024).

Económico (E): En México, la educación superior tiene una estructura de costos que varía significativamente dependiendo del sector institucional, lo cual es un reflejo de las políticas de gratuidad progresiva definidas en la constitución. En este sentido, "las universidades privadas han visto aumentos significativos, comparables a los de otros países como España, aunque mucho menores que los de Estados Unidos; en cambio, las instituciones públicas mantienen precios razonables". (Flores et al., 2023).

De acuerdo con el INEGI (Instituto Nacional de Estadística y Geografía),

> en las universidades públicas, el costo promedio anual de una licenciatura es de aproximadamente $13,824 pesos mexicanos; en cambio, en las privadas, asciende a $48,902 por año. En este sentido, las universidades privadas más caras tienen colegiaturas anuales que oscilan entre los $79,000 (Universidad Tecmilenio) y los $146,900 (Tecnológico de Monterrey), lo que muestra una marcada estratificación económica en la entrada a la educación superior. (INEGI, 2025)

Las cifras de los sueldos por persona y las tasas de desempleo profesional de los graduados universitarios muestran problemas estructurales graves en el mercado laboral mexicano. "En México, los graduados tienen una tasa de empleo del 80.7%, que es menor al promedio de la OCDE (84.1%), lo que indica que una proporción considerable de ellos enfrenta problemas para conseguir un trabajo apropiado a pesar de sus estudios universitarios". (Contreras-Espinoza et al., 2024).

La transición del programa Progresa-Oportunidades-Prospera a las Becas para el Bienestar Benito Juárez experimentó una modificación importante en el sistema de becas mexicano durante el periodo 2018-2024.

> Esta transición significó una transformación radical en la política social educativa, ya que se pasó de un programa que estaba a cargo de la Secretaría responsable de luchar contra la pobreza a uno totalmente gestionado por la Secretaría de Educación Pública mediante la Coordinación Nacional de Becas para el Bienestar Benito Juárez. (Rodríguez Gómez, 2020)

Social (S): La población de estudiantes en México ha crecido de manera notable. "318 programas de salud son ofrecidos por 100 instituciones en Puebla, y se ocupan de 61,432 estudiantes cada año, la mayoría de ellos en centros privados" (García & Pérez, 2023). La percepción social respecto a las carreras universitarias ha ido cambiando, valorando más la formación integral y mostrando un interés cada vez mayor en competencias del siglo XXI y en la educación humanística. Asimismo, los programas del gobierno han procurado tratar la exclusión en términos de empleo y educación de los jóvenes que están en condiciones de vulnerabilidad.

Tecnológico (T): En la educación superior de México, la incorporación de tecnología evidencia progresos importantes. "El 33% de los estudiantes y el 20% de los docentes emplean herramientas de inteligencia artificial, lo que muestra que se están implementando gradualmente tecnologías emergentes" (Chao-Rebolledo & Rivera-Navarro, 2024). Esto demuestra que los estudiantes tienen una mayor aceptación de la aplicación de IA generativa en tareas académicas, resaltando el aumento en la eficiencia y el aprendizaje.

En cuanto al acceso a Internet, muestra diferencias significativas. Por ejemplo, "en Guerrero, en medio de la pandemia, el 27.3% de los alumnos universitarios tenía una buena conexión a Internet, el 43.2% tenía mala conectividad y el 3.4% no contaba con ninguna computadora ni acceso a Internet" (García et al., 2022). Es indiscutible que estas brechas digitales tienen un impacto considerable en la justicia educativa.

Al igual que en la mayoría de los países, en México se aceleró el uso de plataformas digitales durante la pandemia de COVID-19. Dado que las instituciones universitarias han puesto en práctica "modelos híbridos que fusionan la presencialidad con formas virtuales. La Educación 4.0 incluye tecnologías emergentes en el diseño didáctico, pero se choca con retos relacionados con la infraestructura y la formación de los docentes" (Sánchez, 2024).

En este sentido, la tecnología educativa posibilita el uso de modelos predictivos con el fin de detectar a los estudiantes que corren el riesgo de abandonar sus estudios. No obstante, esto necesita marcos regulatorios adecuados para salvaguardar los datos personales y las consideraciones éticas en el tratamiento de la información de los estudiantes.

Ecológicos (E): México afronta diversos riesgos naturales que afectan la infraestructura educativa, debido a su localización geográfica. En este sentido, en el estado de México "las inundaciones tienen un riesgo amplio y generan mayores efectos en la sociedad y el medio ambiente, además; la formación para la administración de riesgos de desastres ha sido incluida poco a poco en ciertos programas educativos" (Paz-Ruiz, 2024).

Además, el acceso y la calidad de la educación se ven impactados en gran medida por las disparidades entre áreas rurales y urbanas. Según Paz-Ruiz, (2024) "la población indígena rural se vio particularmente afectada por la pandemia, lo que empeoró las disparidades educativas existentes". Por lo cual, las instituciones rurales afrontan retos más grandes en términos de conectividad y recursos tecnológicos.

En consecuencia, la educación ambiental ha pasado de visiones ecológicas a puntos de vista más completos. "El currículo de 2022 divide la enseñanza científica ambiental, dando lugar a campos específicos como 'Ética, naturaleza y sociedad', estas posturas ambientales de los estudiantes universitarios reflejan una creciente conciencia, pero necesitan un fortalecimiento en términos de sostenibilidad práctica" (Paz-Ruiz, 2024).

Legal (L): Es esencial que se tenga en cuenta el marco legal de México para la protección de datos personales al aplicar modelos predictivos en la educación superior. "El Instituto Nacional de Transparencia, Acceso a la Información y Protección de Datos Personales (INAI) se encarga de vigilar que las regulaciones de transparencia y de protección de datos se cumplan" (Sánchez & Cruz, 2024).

Por lo que, las instituciones educativas deben considerar las implicaciones legales del manejo de datos estudiantiles, especialmente para análisis predictivos de deserción. La protección de datos sensibles requiere marcos normativos específicos que equilibren la utilidad educativa con los derechos de privacidad.

Implicaciones para el contexto hondureño

En Honduras la educación superior no es obligatoria para toda la población, no obstante, si es inclusiva de la misma manera que se aprecia en México. Al igual que en México los factores económicos tienen una influencia en el porcentaje de deserción en la educación superior en universidades privadas, ya que debido a los costos elevados de las mensualidades y alta matrícula muchas personas dejan de estudiar para enfocarse totalmente al trabajo remunerado y cumplir con las necesidades familiares, aunque los costos de la universidad estatal es significativamente menos a las universidades privadas, muchas personas desertan por diversas situaciones, el bajo rendimiento tiene un alto impacto. Además, los factores sociales también inciden, ya que esa triangulación entre el hogar, trabajo y familia no siempre es bien gestionada por los estudiantes; lo cual se ve reflejado en un bajo rendimiento académico e inclusive el abandono de los estudios. Además, muchas de las personas que se han graduado a nivel universitario no desempeñan funciones asociadas a su carrera profesional; sino que trabajan es aspectos diferentes lo cual está asociado a la alta tasa de desempleo.

Con relación a los aspectos tecnológicos, la mayoría de las universidades cuentan con el uso de plataformas y el uso de IA ha aumentado. No obstante, muchos profesores se resisten al cambio lo cual impacta en el uso inadecuado de estas herramientas por parte de los estudiantes. El deficiente acceso a internet en diversas zonas rurales del país genera que el acceso a la educación se vuelva más complida, sumado a ellos situaciones como desastres naturales que impactan en la misma.

Colombia:

Político (P): La inclusión y la calidad han sido los puntos centrales de las políticas educativas en Colombia. “Las contradicciones entre lo estipulado en las políticas y la realidad que viven las universidades siguen constituyendo grandes desafíos, aunque fortalecer la formación de los profesores y aumentar la inversión pública son prioridades” (Cobo, 2024). Además, la normativa en el acceso a la educación superior en áreas rurales busca luchar contra la inequidad, pero el abismo persiste, perjudicando principalmente a los jóvenes de localidades distantes.

“La acreditación de calidad es responsabilidad del Consejo Nacional de Acreditación (CNA), y ha progresado hacia una perspectiva que tiene como ejes principales la investigación y la internacionalización” (Espinal Ruiz et al., 2020).

Económico (E): Colombia afronta grandes dificultades económicas, ya que numerosas familias consideran la universidad como un lujo costoso e inaccesible. “Las barreras para aquellos que migran a las ciudades, el costo de sostenimiento, la escasa oferta institucional y las disparidades sociales restringen el acceso, particularmente para los estudiantes rurales” (Alarcón, 2025). A pesar de que el Estado fomenta programas de inclusión, créditos subsidiados y becas, estos no siempre benefician a quienes más los necesitan.

Social (S): La desigualdad en el acceso es profunda en Colombia, sobre todo en las universidades públicas, aunque la población estudiantil ha aumentado en las últimas décadas. Los valores familiares y sociales tienen un gran impacto, pues a menudo la decisión de estudiar una carrera universitaria depende de desafiar las normas sociales que ven a la universidad como algo distante para los jóvenes con bajos ingresos. En este sentido, “la inclusión y la igualdad son cuestiones fundamentales en la agenda pública, subrayando lo importante que es contar con políticas que vengan acompañadas de asesoría educativa y orientación vocacional” (Martínez-Garrido & Márquez-Ortíz, 2024).

Tecnológico (T): En Colombia, la digitalización es un desafío y una oportunidad. Debido a que, después de la pandemia, el uso de las TIC en la educación superior aumentó, pero las disparidades en términos de infraestructura y capacitación digital aún persisten, sobre todo en áreas rurales o periféricas a las grandes ciudades. No obstante, “según los estudios de neuroeducación, las capacidades tecnológicas del profesorado y los modelos pedagógicos innovadores están en proceso de fortalecimiento y poseen la capacidad de incrementar el rendimiento académico y la motivación” (Álvarez et al., 2024).

Ecológicos (E): En Colombia, como en México, “las áreas rurales, en particular las que han sido impactadas por desastres naturales o conflictos, tienen más dificultades con la infraestructura educativa; esto repercute en el acceso efectivo a la educación superior y su permanencia” (Verano, 2025). Aunque las políticas públicas tienen como objetivo ser inclusivas, todavía existen desafíos por solucionar en lo que se refiere a infraestructura y movilidad de los estudiantes en los centros urbanos.

Legal (L): La Constitución Nacional de Colombia, en su artículo 67, establece que la educación es un derecho básico. Por lo tanto, la normativa sobre calidad y acceso “se ha ido adecuando a las nuevas circunstancias, aunque todavía queda mucho por mejorar en el marco legal,

particularmente en cuanto al reconocimiento de trayectorias educativas no convencionales y a la protección de datos personales para modelos predictivos" (Ramírez & Castaño, 2024).

Implicaciones para el contexto hondureño

En Colombia desde el Ministerio de Educación Nacional fomentan la digitalización; realizando investigaciones sobre la aplicación de la minería de datos, no obstante, aún se debe trabajar en la protección de los datos, al igual que Honduras, se cuenta con la oportunidad de explorar y aplicara herramientas de análisis de datos que permitan tomar decisiones de forma oportuna en pro de los estudiantes universitarios. Este aspecto de la digitalización es una de las ventajas competitivas que tienen las universidades privadas en Honduras, ya que cuentan con un mejor acceso a la tecnología que las universidades públicas. Por lo cual, pueden implementar diversas herramientas que proporcionen un seguimiento más efectivo de los estudiantes; entre ellos predecir la deserción por diversos factores, como; el bajo rendimiento académico. Colombia apunta a una formación docente más efectiva, no obstante; aún la educación superior es vital como un privilegio caro, caso contrario en Honduras, dónde la matrícula a nivel universitario aumenta o se mantiene, pero de la misma manera la tasa de deserción siendo una de las razones en bajo aprovechamiento académico. En la mayoría de los países la resiente pandemia brindó una oportunidad para reinventarse en brechas digitales, sin embargo; en muchas ocasiones la infraestructura tecnológica y capacitación digital pueden jugar en contra y volverse un factor que afecte la calidad educativa.

España:

Político (P): La coherencia de las políticas nacionales y el impulso europeo son dos aspectos que distinguen a España. Por lo que "el avance de la digitalización en las universidades es bueno debido al apoyo de Europa y del gobierno, así como a la cooperación interregional entre las comunidades autónomas, lo que posibilita una integración tecnológica bastante homogénea" (Vilchis-Torres & Segura-Lazcano, 2025).

La creación de espacios europeos y las reformas significativas, como la LOMLOE, han fomentado la calidad, la internacionalización y los objetivos de sostenibilidad (ODS). "Las políticas también persiguen promover la capacitación continua del profesorado, aunque este proceso demanda una mayor diversidad y una actualización constante" (García & Pastor, 2025).

Con respecto a la acreditación, “los procedimientos y estándares para universidades son supervisados por la Agencia Nacional de Evaluación de la Calidad y Acreditación (ANECA)” (Brunner et al., 2020).

Económico (E): La financiación de la universidad pública en España es mayormente apoyada por el Estado y por los fondos europeos. Esto afirma que la universidad tiene que asegurar la igualdad de acceso, “a pesar de que la demanda en aumento continúa ejerciendo presión sobre las plazas disponibles y los recursos. El 51% de los docentes cree que para garantizar la equidad y el acceso, es fundamental la financiación del gobierno” (Ramírez Díaz, 2024).

A pesar de que los gastos públicos en universidades son sólidos, hay un debate acerca de la eficacia y distribución de recursos en las instituciones para optimizar la retención y graduación de estudiantes. Sin embargo, “las becas y ayudas para estudiantes complementan el sistema público, con un enfoque especial en los grupos vulnerables, fomentando la continuidad de la educación y aliviando las desigualdades económicas” (Montero Caro, 2021).

Social (S): La universidad cumple una función esencial en la movilidad social. En España, se han realizado progresos en términos de democratización y acceso; no obstante, todavía existen obstáculos culturales y socioeconómicos en ciertas áreas y comunidades. Por ejemplo, "la incorporación de estudiantes con discapacidades y de diversas procedencias sociales es un asunto de actualidad. Pero, la competencia digital de los docentes para la inclusión es todavía insuficiente según investigaciones en Andalucía, lo que impide la igualdad real” (Peña, 2009).

Tecnológico (T): La innovación tecnológica y la inversión posicionan a España como un referente en Europa.

> Las universidades se comprometen con las TIC y modelos disruptivos, pero todavía afrontan dificultades y contradicciones en la digitalización de los procedimientos pedagógicos e institucionales. La planificación del futuro de la educación 4.0 se lleva a cabo tomando como referencia modelos institucionales adaptativos y sugerencias de la UNESCO (Castañeda Quintero, 2009).

Ecológico(E): En lo que respecta a la deserción, el impacto ecológico en la educación superior de España no es relevante; sin embargo, la integración de la Agenda 2030 (ODS) y la sostenibilidad son los factores determinantes para modificar los planes de estudio y las gestiones universitarias. En este sentido, “las instituciones han puesto en marcha estrategias para fomentar

transformaciones culturales con un enfoque ambiental, propiciando la sostenibilidad a partir de la formación académica y la administración" (Montero Caro, 2021b).

Legal (L): A pesar de que la descentralización del sistema educativo ha permitido una cierta flexibilidad, el marco legal en España sigue siendo fuerte y centralizado. Por lo tanto, "los requisitos de admisión y la regulación de calidad se ajustan periódicamente a las nuevas exigencias sociales, a las corrientes europeas y a los estándares de diversidad, enfatizando en la transparencia y en la protección del derecho a la educación" (Bonal et al., 2023). En este sentido, la protección de datos personales es un asunto fundamental en la administración universitaria, especialmente para modelos predictivos que se fundamentan en el *big data* académico.

Implicaciones para el contexto hondureño

España muestra un avance significativo en digitalización, volviéndose uno de los referentes en inversión e innovación tecnológica. Además, existe mayor oportunidad de acceder a becas, créditos subsidiados y programas de inclusión, todo ello con el fin de mejorar la retención y aumentar el de graduación. Aunque; no siempre llegan a quienes más lo necesitan. Sin embargo, muchos países latinoamericanos se benefician de estas becas entre ellos Honduras, donde existe la oportunidad de optar a becas en España bajo el cumplimiento de ciertos requisitos.

La inclusión con relación a la educación, en Honduras ha demostrado un notorio avance, no obstante, en España es un tema que se encuentra aún sobre la mesa, lo cual impacta en el acceso a las becas por parte de sus ciudadanos. Al igual que en Honduras, la inserción de los Objetivos de Desarrollo sostenible (ODS), marcan una pauta en las reformas curriculares.

El marco legal de España es flexible y se adapta nuevas demandas, enfocado en la transparencia y protección de derechos de la educación, en Honduras, esto sería de mucho beneficio ya que al lograr mayor flexibilidad en los aspectos legales se puede aspirar a contar con una mejor calidad educativa que no solo se adapte a las tendencias tecnológicas, sino que abarque todos los aspectos relacionados con brindar una mejor y accesible experiencia universitaria.

2.2 ANÁLISIS DEL MICROENTORNO

Esta investigación utiliza las Cinco Fuerzas de Porter para analizar el sector educativo universitario en El Salvador, Guatemala y Costa Rica, con la finalidad de proporcionar un contexto regional sólido y contrastarlo con Honduras.

El Salvador:

Rivalidad entre competidores existentes: En El Salvador hay una fuerte competencia entre entidades públicas y privadas, como la Universidad de El Salvador y la Universidad Tecnológica de El Salvador o la Universidad Católica. No obstante, “las universidades públicas tienen una demanda elevada porque son más asequibles desde el punto de vista económico, mientras que las privadas brindan una mayor especialización y flexibilidad en cuanto a horarios” (Universidad de El Salvador, 2022).

Amenaza de nuevos entrantes: En El Salvador, las barreras regulatorias son moderadas; "el establecimiento de nuevas universidades, en particular virtuales, ha sido posible gracias a la digitalización pospandémica” (Rivas, 2023). La difusión de instituciones privadas pequeñas continúa, a pesar del mejoramiento de la acreditación estatal en los años recientes. La emergencia sanitaria global contribuyó a que aumente el número de universidades privadas.

Poder de negociación de los proveedores: Dentro de los proveedores clave se encuentran docentes especializados, recursos tecnológicos y plataformas de aprendizaje; sin embargo, su poder es limitado. En este sentido, "las universidades públicas establecen alianzas con organizaciones internacionales para elevar su calidad, a la vez que las privadas intentan contratar personal académico con experiencia en educación virtual y perfiles técnicos” (Ruiz, 2025). Asimismo, los costos relacionados con esta captación representan un reto en términos financieros, y la falta de talento especializado puede restringir la habilidad de mantener innovaciones educativas que contribuyan a detectar con anticipación y prevenir eficazmente el abandono escolar de estudiantes en riesgo.

Poder de negociación de los clientes (estudiantes): La expansión del acceso a la información posibilita que los estudiantes y sus familias se conviertan en clientes influyentes, que no solo exigen educación académica, sino también servicios completos de soporte, como tutorías personalizadas, herramientas digitales y alternativas de financiación. “Los programas de becas y ayudas del gobierno complementan la capacidad de elección de los estudiantes, aunque continúan existiendo restricciones socioeconómicas que obstaculizan la continuidad académica frente a reprobaciones repetidas” (LATAM, 2025) La personalización del acompañamiento basado en modelos predictivos de riesgo se convierte en un medio distintivo dentro de una competencia elevada.

Amenaza de productos o servicios sustitutos: Para los estudiantes interesados en escapar de la inflexibilidad del modelo universitario clásico, el avance de la educación técnica y el aumento de cursos breves brindan opciones interesantes. Por lo tanto, “las certificaciones digitales internacionales y las plataformas educativas masivas empiezan a consolidarse como alternativas que facilitan la adquisición de habilidades esenciales, lo cual es un reto para las universidades tradicionales, en cuanto a mantener a los estudiantes con problemas académicos” (Grigorio & Pereira, 2025).

Guatemala

Rivalidad entre competidores existentes: Guatemala tiene una red universitaria variada, que incluye universidades públicas y privadas destacadas, como la Universidad de San Carlos, la Universidad Galileo y la Universidad del Valle de Guatemala. "En cuanto a la competencia, se basa en la calidad, la accesibilidad y los programas innovadores; sin embargo, hay diferencias notables en términos de cobertura e infraestructura, especialmente en áreas rurales” (López et al., 2024). Las entidades educativas se centran en impedir que los estudiantes de primer año abandonen sus estudios, a través de un seguimiento académico estructurado que utiliza análisis predictivos de riesgo y tecnologías.

Amenaza de nuevos entrantes: Según Beltrán, (2024) “Se ha incrementado la creación de universidades privadas y la ampliación de modalidades semipresenciales o virtuales, gracias a reglamentos universitarios permisivos; sin embargo, la necesidad de acreditación y de calidad ralentiza el ingreso desordenado”. Si bien la entrada de nuevos actores es facilitada por los avances tecnológicos, para que se consoliden deben desarrollar servicios especializados para ayudar a los estudiantes en riesgo de abandono, lo cual requiere inversiones significativas.

Poder de negociación de los proveedores: La capacidad de negociación de los docentes capacitados para la educación virtual y el análisis de datos aumenta debido a la escasez de estos. Por lo que “las universidades que consiguen formar alianzas estratégicas con proveedores de tecnología y organizaciones educativas a nivel global hacen más fácil la transformación digital y el desarrollo de modelos predictivos para identificar a los estudiantes que tienen problemas desde etapas tempranas” (Zavala et al., 2025).

Poder de negociación de los clientes (estudiantes): Los estudiantes de Guatemala valoran la propuesta académica que incluye programas de empleabilidad asegurada, becas y sistemas de respaldo individualizados. El aumento en el acceso a la información y la movilización de los estudiantes hacen que el poder se fortalezca. Como respuesta, “las universidades han implementado tácticas para reducir la deserción que se basan en el análisis de los patrones de rendimiento, el seguimiento individual y estrategias motivacionales” (López, 2025).

Amenaza de productos o servicios sustitutos: Al igual que en El Salvador, “los programas técnicos y los cursos en línea, junto a la posibilidad de obtener certificaciones profesionales, captan a estudiantes que buscan acceder al mercado laboral de forma rápida, representando una alternativa válida ante la educación universitaria convencional” (Contreras López et al., 2022), esto ocurre sobre todo entre aquellos estudiantes con un riesgo más alto de dejar los estudios formales.

Costa Rica

Rivalidad entre competidores existentes: El sistema universitario de Costa Rica lo constituyen universidades públicas como el Instituto Tecnológico de Costa Rica y la Universidad de Costa Rica, así como un sector privado en expansión. Es por ello que su rivalidad es intensa tanto entre las universidades públicas como las que pertenecen al sector privado "los ámbitos de competencia incluyen la calidad del currículo, la investigación aplicada y el aumento de las oportunidades laborales para los estudiantes"(Casillas Alvarado et al., 2021b). Para reducir la deserción, en particular durante los primeros años con un alto índice de reprobación, se utilizan modelos predictivos de riesgo y protocolos de seguimiento; además de competir por acreditaciones, prestigio y sobre todo en las estrategias de retención.

Amenaza de nuevos entrantes: La entrada de nuevas universidades “es restringida por las rigurosas regulaciones en cuanto a acreditación y calidad, aunque el crecimiento de tecnologías digitales incrementa la variedad de programas virtuales y flexibles”(Mejía González et al., 2022). No obstante, el efecto en la retención de estudiantes con un rendimiento académico bajo sigue siendo un eje clave de la política educativa en Costa Rica, ya que las nuevas modalidades hacen más fácil aumentar el acceso. Además, de la competencia por parte de universidades internacionales a través de la modalidad virtual, lo impulsa a las universidades locales a modernizar sus enfoques del proceso de enseñanza-aprendizaje.

Poder de negociación de los proveedores: Actualmente en Costa Rica se ha optado por migrar sus cursos de la modalidad presencial a la virtual, para lo cual "requieren personal altamente calificado, además de un diseño instruccional y acceso a materiales de calidad para marcar la diferencia en la experiencia estudiantil y en la retención"(Villalobos et al., 2023), además; a la hora de buscar plataformas para el análisis de datos, tecnologías avanzadas y personal académico especializado, el poder de los proveedores se fortalece, lo cual es esencial para individualizar la atención a los estudiantes en riesgo. En este sentido, "para una educación de excelencia y para evitar la deserción temprana, son necesarias las estrategias de cooperación internacional y apoyo tecnológico" (Cantero-Acosta & Bolaños-Ortíz, 2020).

Poder de negociación de los clientes (estudiantes): Los estudiantes costarricenses se han vuelto más exigentes, demandando soporte académico, sistemas más flexibles, así como recursos para el apoyo académico, asesoría individualizada y herramientas virtuales. Es por ello que "el acceso a becas privadas y estatales estimula la matrícula, pero en el momento de escoger una universidad se vuelve más fuerte la exigencia de calidad y resultados concretos"(Huitrón, 2020).

Amenaza de productos o servicios sustitutos: Las opciones de cursos masivos abiertos (MOOC), programas cortos, certificaciones a nivel internacional y educación técnica "constituyen alternativas a la universidad tradicional, especialmente para aquellos que desean una formación específica y rápida con una menor inversión de tiempo y dinero"(Achoy Sánchez & Jiménez Segura, 2023). En este sentido el reto es innovar sus modelos de acompañamiento y persistencia estudiantil.

Honduras

Rivalidad entre competidores existentes: La Universidad Nacional Autónoma de Honduras (UNAH) tiene el control del sector público junto con universidades privadas establecidas como UNITEC, Universidad Católica de Honduras y UTH. Según Marchesi & Hernández, (2019) "la competencia se fundamenta en la innovación tecnológica, los servicios académicos individualizados y la especialización, que abarcan el seguimiento a estudiantes con riesgo de deserción". En este sentido, una de las tácticas que están surgiendo para aumentar la retención es implementar modelos que generen una empleabilidad real en los estudiantes, además de un análisis educativo de acompañamiento y seguimiento más automatizado.

Amenaza de nuevos entrantes: La competencia incrementó pese a que las barreras regulatorias son moderadas, debido a "la rápida implementación de modalidad virtual lo que ha facilitado la llegada de nuevos competidores y la expansión del mercado educativo" (Montero Caro, 2021c). El reto es preservar la calidad y prevenir que el crecimiento afecte la sostenibilidad y el apoyo apropiado a los estudiantes en situaciones de vulnerabilidad, con enfoques más innovadores y una efectiva atención estudiantil como un elemento diferenciador.

Poder de negociación de los proveedores: Los proveedores académicos y tecnológicos tienen un poder significativo, debido a que hay escasez de personas con la habilidad "para poner en práctica sistemas de educación analítica avanzada que hagan posible localizar a los estudiantes que están en riesgo de abandonar y que han reprobado clases"(Espina, 2022). Ya que este recurso humano es limitado y costoso.

Poder de negociación de los clientes (estudiantes): El estudiantado hondureño es exigente "en cuanto a opciones flexibles, acceso a becas, calidad académica y los servicios de orientación personalizados, así como un soporte académico automatizado" (Sosa, 2022). Por consiguiente, su poder de elección es mayor gracias al incremento de la competencia entre universidades públicas y privadas. Por lo que los sistemas de becas y apoyos aumentan su potencial para elegir y requieren modelos eficaces que ayuden a prevenir la deserción.

Amenaza de productos o servicios sustitutos: En un mercado educativo que cambia constantemente y se encuentra vinculado con las exigencias laborales actuales, en este sentido "la formación técnica, los programas de duración breve y los certificados digitales tienen un papel cada vez más importante como opciones válidas" (Yagual et al., 2022). Sobre todo, para estudiantes con bajo rendimiento académico y por ende en riesgo de deserción, esta alternativa es muy atractiva; esto obliga a las universidades a innovar en métodos, contenido y en servicios de retención.

2.3 CONCEPTUALIZACION

Deserción Estudiantil

Se define como el abandono definitivo o parcial de los programas académicos antes de culminar el plan de estudios , este tipo de casos en la educación superior afectan la eficiencia

terminal y representan un reto crítico, ciertos factores asociados incluyen las condiciones académicas, institucionales, socioeconómicas, personales, tecnológicos y contextuales (Pereira Santana & Vidal Cortez, 2020).

La deserción estudiantil se medirá como la cantidad de estudiantes que no culminan su carrera universitaria, identificados a través de los sistemas de matrícula con una la baja total por más de un año; o por el abandono del estudiante en dos periodos consecutivos, lo cual será denominado un desertor temprano (Miño de Gauto, 2021).

Seguimiento Académico Tradicional

Se basa en las practicas administrativas ya sea como el control de asistencia, las revisiones periódicas de puntajes, entrevistas personales y los programas de tutorías, aunque estas estrategias nos permiten identificar a estudiantes en alto riesgo educativo, son más reactivas y no logran una identificación temprana de la deserción estudiantil (Terraza-Beleño, W., 2019).

El seguimiento académico constituye las estrategias de acompañamiento que se le proporcionan al estudiante, tales como; controles de asistencia, revisiones de su historial académico, realización de entrevistas para seguimiento, identificación de necesidad de tutorías, contabilizando la frecuencia por periodo del requerimiento de estas por parte de los estudiantes.

Minería de Datos

Es el proceso de análisis y exploración sistemática de enormes cantidades de información para descubrir patrones, relaciones y tendencias, en la educación superior la minería de datos es empleada para el análisis de bases de datos académicas, socioeconómicas y de conducta con el objetivo de anticipar los comportamientos de deserción (Cristobal Romero & Ventura, 2020).

La minería de datos es medida a través de la aplicación de técnicas y algoritmos predictivos que permitan la identificación de patrones y tendencias mediante el análisis de base de datos académicos de los estudiantes con el propósito de anticipar el riesgo de deserción por parte de los estudiantes.

Modelos Predictivos

Son herramientas estadísticas e informáticas que con el uso de datos históricos y actuales nos permiten estimar la probabilidad de ocurrencia de un evento futuro, en la educación superior los modelos predictivos nos han demostrado que son efectivos para anticipar riesgos de abandono

con el uso de técnicas como la regresión logística, arboles de decisión, mediante algoritmos de aprendizaje supervisado y no supervisado (Villar & De Andrade, 2024).

La aplicación de modelos predictivos permite la estimación de riesgo de deserción estudiantil medidos a través del árbol de decisiones y la regresión logística para el procesamiento de datos de rendimiento académico que determinen la probabilidad de abandono.

Análisis Comparativo

Esta metodología nos permite evaluar de manera critica las similitudes, diferencias, ventajas, desventajas y limitaciones de los enfoques alternativos, es por esto que el análisis comparativo contrasta la minería de datos con seguimiento académico tradicional, nos ayuda a identificar cuál de los dos enfoques resulta más eficaz para la prevención en la deserción estudiantil (Parra-Sánchez et al., 2023).

El análisis comparativo evalúa la eficacia de la implementación de minería de datos versus métodos tradicionales de seguimiento académico para anticipar el riesgo de deserción estudiantil; contrastando similitudes, diferencias y beneficios, con base en indicadores de precisión predictiva, tiempo de detección entre otros.

Factores de Riesgo

Los factores son variantes o condiciones que pueden llegar a incrementar o afecta la probabilidad de abandono de los estudios superiores, estos factores se pueden clasificar en categorías como académicos (bajo rendimiento), institucionales (falta de programas de tutoría), socioeconómicos (tener que trabajar), tecnológicos (el desconocimiento digital), personales (la falta de motivación) y contextuales (distancias y transporte) (Garrido Silva & Pajuelo Diaz, 2023).

Los factores de riesgo determinan las condiciones que incrementan la probabilidad de abandono académico por parte de los estudiantes, lo que permite identificar y cuantificar los factores que inciden en la deserción estudiantil.

2.4 TEORÍAS DE SUSTENTO

2.4.1 BASES TEÓRICAS

El rendimiento académico y la deserción estudiantil ha sido analizado múltiples veces en

varias investigaciones con enfoques teóricos, los cuales buscan dar una explicación a los factores que generan la deserción estudiantil en las universidades, esto es algo que se observa desde los años 70s con las teorías como la de Vincent Tinto, que explica la integración académica y social, además de los modelos de retención de *bean*, demuestran que la estadía del estudiante depende de variables personales, institucionales y socioeconómicas.

Recientemente estas bases se fortalecieron con la introducción del aprendizaje automático y el análisis predictivo, el cual permite identificar patrones de deserción con mayor eficacia y ayuda a disminuir el riesgo. Un ejemplo podría ser Villegas-Ch et al., (2023) que muestra que el uso de *machine learning* contribuye a la mejora de los índices de retención en la educación superior al detectar a los estudiantes en posible riesgo de deserción, también González-Nucamendi et al., (2023) plantea que los modelos predictivos interpretables ayuda a la identificación de estudiantes vulnerables, lo que mejora las invenciones académicas.

Los algoritmos de *machine learning* permiten realizar una predicción más exacta sobre estudiantes propensos a desertar con base en el análisis de su desempeño académico, es una forma más efectiva de obtener esta información de manera más rápida; lo que permite brindar un seguimiento más oportuno a los estudiantes. En esta investigación se utilizará algoritmos predictivos para determinar de forma automatizada los patrones que permiten anticipar el riesgo de deserción estudiantil.

2.4.2 METODOLOGÍAS DESARROLLADAS

Los análisis predictivos de los riesgos de la deserción en las universidades se apoyan en los algoritmos de clasificación, predicción y agrupación, que nos ayudan a modelas variables que se relacionan con el historial académico, la parte socioeconómica y los patrones de comportamientos de los estudiantes.

- **Arboles de decisión:** los cuales se utilizarán para segmentar poblaciones de estudiantes que den función a sus características, estas ofreciéndonos reglas claras y fáciles de interpretar (Hoyos Osorio & Daza Santacoloma, 2023).
- **Regresión Logística:** se utiliza para determinar la probabilidad en la que un estudiante abandone los estudios universitarios, mostrando altos niveles de precisión. (Aguilar López et al., 2024).

- **Random Forest:** esta técnica mejora los índices de predicción debido a que combinas múltiples arboles de decisión, lo que nos ayudaría reduciendo la tasa de riesgo del sobreajuste. (Aguilar López et al., 2024).

Esta investigación integra metodologías basadas en algoritmos para analizar el rendimiento académico de los estudiantes y determinar patrones que anticipen de forma acertada el riesgo de deserción estudiantil. Aplicando el árbol de decisiones para segmentar a los estudiantes con base en sus características académicas, ya que permite identificar los grupos con mayor riesgo mediante la comprensión de los factores que influyen en ello. Por su parte la regresión logística calcula la probabilidad de que un estudiante abandone sus estudios universitarios, ya que este enfoque estadístico es altamente preciso en clasificación binaria.

2.4.3 INSTRUMENTOS UTILIZADOS

- **Bases de datos institucionales:** los cuales serían historial académico, calificaciones, asistencia, variables demográficas.
- **Plataformas de minería de datos y *machine learning*:** como WEKA y Python *Scikit-learn*, utilizadas para implementar modelos de predicción.
- **Los algoritmos de clasificación, predicción y agrupación:** estos algoritmos de minería de datos, se aplicaran a los contextos educativos que se utiliza en la clasificación de los estudiantes según su probabilidad de graduación o deserción, predecir su rendimiento a futuro y agrupar los estudiantes con características similares, según (González-Nucamendi et al., 2023), estos enfoques son efectivos debido a que permiten reconocer patrones ocultos en datos institucionales, ya sea historial académico, asistencia y factores socioeconómicos.
- **Árboles de decisiones:** utilizaremos árboles de decisiones debido a que son los más utilizados en la educación superior debido a su interpretabilidad y su fácil implementación. Utilizando los árboles de decisiones se dividirá a los estudiantes en grupos basados en reglas, como promedio de clases reprobadas o promedios bajos, así como las clases retiradas o sin derecho. (Hoyos Osorio & Daza Santacoloma, 2023) mencionan que los árboles de decisiones destacan por la capacidad de generar reglas sencillas que nos facilite identificar a estudiantes en riesgo durante las etapas tempranas.

- **Regresión Logística:** la regresión logística es uno de los métodos más precisos para la predicción de los riesgos de deserción, se utiliza para realizar cálculos en cuanto a la probabilidad de que los estudiantes abandonen sus estudios en funciones variables independientes, ya sea el rendimiento académico y la escasez de recursos económicos.

Los instrumentos aplicados en la investigación han sido seleccionados con finalidad de recopilar organizar y procesar la información que permita identificar mediante un análisis predictivo a los estudiantes que se encuentran en riesgo de deserción.

Gracias al empleo de las bases de datos institucionales, con la información sobre el historial académico, calificaciones, asistencia y estado de las asignaturas, además; de las herramientas de minería de datos que permitan automatizar los procesos de clasificación y predicción con base en su trayectoria académica. El árbol de decisiones como una técnica para segmentar a los estudiantes de acuerdo con la cantidad de clases reprobadas, su bajo promedio, clases retiradas y clases sin derecho; complementando con la regresión logística que es un instrumento de predicción en función de las variables, todo ello; con el propósito de asegurar la validez de la predicción e información confiable para tomar decisiones orientadas a la retención de estudiantes.

2.5 ANÁLISIS DE LAS METODOLOGÍAS

En el estudio de la deserción estudiantil en las universidades se ha visto una evolución en los últimos años gracias a la implementación e ideación de técnicas de minería de datos y modelos predictivos, estas metodologías permiten identificar ciertos patrones ocultos en grandes cantidades de información académica, lo cual nos brinda una ventaja frente a enfoques tradicionales, debido a la agilidad y eficiencia de los enfoques modernos y su práctica implementación, los enfoques modernos destacan por su agilidad, capacidad preventiva y aplicabilidad practica (González-Nucamendi et al., 2023).

Los árboles de decisión son una herramienta ampliamente utilizada debido a su interpretabilidad y por la sencillez en su implementación, con estos algoritmos podemos dividir a los estudiantes universitarios en grupos mediante reglas claras y concisas, ya sea por cantidad de reprobados, promedio de baja, lo que facilite la detección temprana de perfiles en riesgo, su fortaleza es que ofrece explicaciones comprensibles para docentes y autoridades universitarias,

debido a esto se presentan limitaciones relacionadas al sobreajuste, por lo que se recomienda completar su uso con técnicas de validación cruzada o con el robustecimiento mediante algoritmos (Hoyos Osorio & Daza Santacoloma, 2023).

Como segundo punto, la regresión logística nos ha demostrado que es una de las metodologías más certeras para estimar la probabilidad de deserción, este modelo permite relacionar variables independientes con la variable dependiente, su ventaja radica en la capacidad y beneficio de calcular probabilidades con un nivel de confianza estadístico alto, no obstante su debilidad es la dependencia de supuestos lineales que en ciertas ocasiones, no capturan la complejidad de los factores de los estudiantes universitarios (Aguilar López et al., 2024).

Una metodología que se ha consolidado como de alto desempeño es la metodología de Random Forest ya que nos permite combinar múltiples arboles de decisión, este enfoque no solo mejora la precisión en la predicción, sino que también reduce la varianza y evita el sobreajuste que presentan los árboles simples, la limitación es su complejidad durante la interpretación de los resultados, ya que el modelo funciona como una "caja negra" para los usuarios que no están especializados (Villegas-Ch et al., 2023).

La regresión logística nos ha demostrado que es una de las metodologías más certeras para estimar las probabilidades de deserción, es uno de los modelos que nos permiten relacional un conjunto de variables independientes con una variable binaria que simboliza desertor o no desertor, su principal ventaja es calcular probabilidades con un nivel de confianza estadístico alto, la cual facilita la toma de decisiones basadas en evidencias (Aguilar López et al., 2024).

A diferencia de los métodos tradicionales de seguimiento académico, los modelos predictivos nos brindan ventajas significativas, aunque el seguimiento clásico suele ser más reactivo a la hora de identificar problemas cuando estos ya se han manifestado, las metodologías de minería de datos permiten una intervención temprana y preventiva, que se orienta a una mejor retención de estudiantes (González-Nucamendi et al., 2023).

En conclusión, la utilización combinada de los modelos predictivos ofrece a las universidades la posibilidad de anticipar riesgo de deserción con mayor precisión y confiabilidad, mientras los árboles de decisión aportan claridad y facilidad de interpretación, la regresión logística nos permite cuantificar probabilidades de forma sencilla y el uso de *random forest* proporciona un alto desempeño predictivo en escenarios complejos.

2.6 ANTECEDENTES DE LAS METODOLOGÍAS

El uso o la implementación de las metodologías de la minería de datos y modelos predictivos para el análisis de la deserción de los estudiantes universitarios, no es reciente, sin embargo, durante los últimos años se ha consolidado debido al rápido avance de técnicas de aprendizaje automatizado y el incremento de disponibilidad de bases de datos académicas.

Uno de los modelos que fueron pioneros en el campo de enfoques aplicados son los modelos estadísticos clásicos, entre los cuales destaca la regresión logística, que ha sido ampliamente utilizada para estimar ciertas probabilidades de abandono en función a variables socioeconómicas y académicas, según estudios recientes han confirmado su utilidad en universidades de Latinoamérica, al identificar varios factores de riesgos asociados al pobre rendimiento, escasa participación de tutorías y deficiencia en la comunicación con docentes (Aguilar Lopez et al., 2024).

Con la evolución de la minería de datos los árboles de decisión comenzaron a emplearse para poder segmentar a los estudiantes en función de sus características o cualidades, ya sea como número de asignatura reprobadas, asistencia o historial académico. Su atractivo radica en interpretabilidad, lo que nos permite comprender o analizar de manera sencillas las causas que generan la permanencia o deserción (Hoyos Osorio & Daza Santacoloma, 2023).

Recientemente, el reciente desarrollo de los algoritmos avanzados como Random Forest, *Support* Vector Machines y Redes Neuronales Artificiales ha visto un incremento en la precisión de los modelos, estas metodologías han demostrado un desempeño superior en predicción de deserción, siempre con ciertas limitaciones en la explicación de resultados y la dificultad en la adopción para los contextos de educación donde la transparencia es crucial (González-Nucamendi et al., 2023).

2.7 METODOLOGÍAS, ENFOQUES, MÉTODOS Y DISEÑOS

La elección de metodologías, métodos y diseños de investigación es crucial y depende de las demandas particulares del contexto de investigación. Cada metodología aporta una estructura única y está diseñada para problemas específicos. Para el presente trabajo, que busca analizar comparativamente modelos predictivos de deserción estudiantil, se adopta un enfoque cuantitativo, un método de análisis comparativo y un diseño no experimental de tipo predictivo.

2.7.1 ENFOQUE

La investigación se enfocará en el método cuantitativo. Este método se distingue por la recopilación y el análisis de datos numéricos para verificar hipótesis y determinar patrones conductuales. En la investigación, los siguientes puntos demuestran esto.

El objetivo general es: Evaluar la efectividad de la aplicación de técnicas de minería de datos y *learning analytics* para la predicción y anticipación del riesgo de deserción, mediante indicadores de precisión y capacidad de anticipación. La evaluación de la efectividad y la precisión, y la cuantificación de su desempeño, son intrínsecamente cuantitativas.

Se busca Determinar la precisión predictiva y la capacidad de anticipación de los modelos de árboles de decisión y regresión logística, cuantificando su desempeño. La cuantificación es un pilar crucial de la metodología cuantitativa.

La investigación se basa en el análisis de grandes volúmenes de información y datos históricos de bases de datos académicos de los estudiantes, utilizando algoritmos de clasificación, predicción y agrupación para identificar patrones ocultos.

2.7.2 MÉTODO

El método central de la investigación es el análisis comparativo. Tal metodología permite evaluar de manera crítica las similitudes, diferencias, ventajas, desventajas y limitaciones de los enfoques alternativos. Con esto, se contrastará la efectividad de las técnicas de minería de datos y *learning analytics* con el seguimiento académico tradicional. Se pretende determinar cuál de los dos enfoques es más eficaz para la prevención en la deserción estudiantil.

2.7.3 DISEÑO

El diseño específico de la investigación es no experimental de tipo predictivo.

No experimental: La investigación no manipula intencionalmente variables; en su lugar, se observan y analizan fenómenos ya existentes, como el desempeño académico y los datos históricos de los estudiantes. Los datos institucionales disponibles son la base para construir y validar los modelos.

Predictivo: El objetivo principal es anticipar el riesgo académico y "redecir posibles escenarios de riesgo académico utilizando modelos predictivos. Estos modelos, como los árboles de decisión y la regresión logística, permiten estimar la probabilidad de ocurrencia de un evento futuro, en este caso, la deserción estudiantil.

Justificación Epistemológica

Por las razones que se indican a continuación, la combinación de un enfoque cuantitativo, un método de análisis comparativo y un diseño predictivo es el más adecuado para describir el fenómeno de la deserción estudiantil y dar respuestas a las preguntas de la investigación:

Objetividad y mediciones exactas: El método cuantitativo posibilita la evaluación objetiva de la exactitud y eficacia de los modelos. Para abordar un asunto tan crucial como la deserción estudiantil, que tiene un impacto tanto en la calidad de la educación como en la eficacia de las instituciones, es fundamental disponer de pruebas numéricas robustas para tomar decisiones fundamentadas en datos.

Identificación de Patrones Ocultos: La minería de datos, fundamental en el enfoque predictivo, es el proceso de análisis y exploración sistemática de enormes cantidades de información para descubrir patrones, relaciones y tendencias. Esta capacidad es crucial para identificar patrones que podrían prever el bajo desempeño o el riesgo de deserción de los estudiantes, algo que los métodos tradicionales no logran con suficiente antelación.

Prevenir e intervenir a tiempo: El diseño predictivo posibilita una intervención temprana y preventiva, en contraposición con el seguimiento académico tradicional, que es más reactivo. Esta es una necesidad urgente para las universidades que desean eludir a tiempo los peligros que afrontan los estudiantes y mejorar los programas de apoyo.

Evaluación de la eficiencia de las estrategias: Es adecuado el método comparativo para cotejar cuán eficaces son los árboles de decisión y la regresión logística, por ejemplo, en comparación con los métodos tradicionales de seguimiento. De esta manera se responde de forma directa a la cuestión sobre la eficacia de las técnicas de minería de datos. Esto hace posible que se propongan recomendaciones específicas, basadas en evidencias, para mejorar las estrategias de acompañamiento estudiantil.

Solidez y fundamentación teórica: La aplicación de algoritmos, tales como la regresión logística y los árboles de decisión, se sustenta en bases teóricas sólidas que han sido fortalecidas gracias al análisis predictivo y el aprendizaje automático. Esto ha permitido un reconocimiento más eficaz de los patrones de deserción.

2.8 ANÁLISIS CRÍTICO DE LAS METODOLOGÍAS

La elección de un enfoque cuantitativo, un método de análisis comparativo y un diseño no experimental de tipo predictivo no responde a una mera receta metodológica, sino a una evaluación crítica de las demandas particulares del contexto de investigación y la naturaleza del fenómeno de la deserción estudiantil. Cada metodología aporta una estructura única y está diseñada para problemas específicos, y para el estudio de la deserción en la educación superior, donde se manejan grandes volúmenes de información y la meta es la intervención temprana y preventiva, este camino es el más pertinente.

Defensa de la Idoneidad Metodológica y Reconocimiento de Limitaciones

La elección de un foque cuantitativo, un método de análisis comparativo y un diseño experimental de tipo predictivo no responden a una simple receta metodológica, si no responden a una evaluación de las demandas particulares del contexto de investigación y de la naturaleza del fenómeno de la deserción estudiantil, es por eso que cada deserción metodológica responde a la necesidad de estudiar un fenómeno complejo que afecta a las instituciones de educación superior.

Enfoque Cuantitativo: Es apropiado para analizar la eficacia de emplear técnicas de minería de datos y análisis de aprendizaje en el pronóstico y anticipación del riesgo de deserción, utilizando indicadores de predicción precisa y capacidad anticipatoria. Para calcular la capacidad de previsión y la precisión predictiva es necesario realizar un sólido análisis numérico (Creswell, J. W., & Creswell, J. D., 2018).

Aunque los métodos convencionales de seguimiento académico cumplen su función, "no son suficientes para prevenir a tiempo los riesgos que afrontan los estudiantes", ya que son más reactivos que preventivos. Utilizando la minería de datos, el enfoque cuantitativo facilita "el descubrimiento de factores de riesgo y patrones ocultos asociados a un desempeño deficiente", que en otras circunstancias no serían detectados.

Método de Análisis Comparativo: Este método es el eje central del estudio, ya que posibilita la evaluación crítica de las similitudes y diferencias, así como de los beneficios, inconvenientes y restricciones de las distintas perspectivas (González-Nucamendi et al., 2023). Es

fundamental comparar los modelos predictivos (regresión logística y árboles de decisión) con el seguimiento académico tradicional para determinar cuál de las dos metodologías es más efectiva para prevenir que los estudiantes abandonen sus estudios.

Diseño No Experimental Predictivo: Se justifica plenamente dado que la investigación no busca manipular variables, sino observar y analizar datos históricos para anticipar el riesgo académico y predecir posibles escenarios. La capacidad de estimar la probabilidad de ocurrencia de un evento futuro como la deserción estudiantil, a partir de datos históricos y actuales, es la esencia de este diseño.

Limitaciones Reconocidas: A pesar de su robustez, los modelos predictivos y las técnicas de minería de datos tienen limitaciones. Los árboles de decisión, si bien son interpretables y de fácil implementación, pueden presentar "imitaciones relacionadas al sobreajuste, lo que requiere técnicas de validación cruzada o con el robustecimiento mediante algoritmos. A pesar de que la regresión logística es precisa para calcular la probabilidad de deserción, se basa en supuestos lineales que a veces no logran reflejar la complejidad de los factores relacionados con los estudiantes universitarios (Hoyos Osorio & Daza Santacoloma, 2023).

Además, algoritmos más sofisticados como Random Forest, a pesar de que aumentan la precisión, pueden resultar en una caja negra y mostrar complejidad al interpretar los resultados para aquellos usuarios que no son expertos. Estas restricciones se tratarán a través de la implementación de criterios rigurosos y la combinación de diferentes técnicas (Villegas-Ch et al., 2023).

Unión Explícita entre Teoría y Práctica Metodológica

La investigación se sustenta en bases teóricas que, desde las teorías de retención estudiantil, han evolucionado con la introducción del aprendizaje automático y el análisis predictivo. Conceptos clave como la minería de datos educativa que busca descubrir de manera oportuna un cambio en el comportamiento vinculado a aspectos académicos que puedan predecir una inminente deserción o abandono escolar y la analítica del aprendizaje medición, recopilación, análisis e informe de datos sobre los estudiantes y sus contextos, con el fin de comprender y optimizar el

aprendizaje se materializan directamente en la práctica metodológica (Cristobal Romero & Ventura, 2020).

La aplicación de algoritmos de clasificación, predicción y agrupación específicamente los árboles de decisión y la regresión logística sobre datos históricos de bases de datos académicos de los estudiantes, es la operacionalización de estas teorías y conceptos. Estos modelos permiten identificar patrones que podrían prever el bajo desempeño o el riesgo de deserción, trasladando el marco conceptual al análisis empírico. La minería de datos, entonces, no es solo una herramienta, sino la vía para manejar o transformar grandes volúmenes de información y arrojar nueva luz sobre estudiantes y los entornos educativos, conectando directamente la teoría con la capacidad de generar conocimiento práctico.

Criterios de Rigor y Calidad

Para asegurar la validez y confiabilidad de los resultados en este estudio cuantitativo, se aplicarán los siguientes criterios:

Precisión y Validez Predictiva: La investigación se centrará en cuantificando su desempeño mediante indicadores de precisión y capacidad de anticipación. Esto implica la evaluación de la capacidad de los modelos para clasificar correctamente a los estudiantes en riesgo y predecir la deserción antes de que ocurra. Se manejarán métricas estándar en minería de datos, como la exactitud, precisión, exhaustividad, y el área bajo la curva, para cotejar objetivamente los modelos (Hassan et al., 2024).

Confiabilidad y Robustez: Se efectuarán técnicas de validación cruzada para afirmar que los modelos no estén sobre ajustados y que sus resultados sean sólidos y generalizables a nuevas poblaciones de datos. Para los árboles de decisión, se buscará completar su uso con técnicas de validación cruzada o con el robustecimiento mediante algoritmos, también contribuye a reducir el sobreajuste. Herramientas como R con librerías permiten una validación cruzada rigurosa.

Calidad y Seguridad de los Datos: Se garantizará la calidad de los grandes volúmenes de información provenientes de las bases de datos institucionales. Se seguirán estándares internacionales para asegurar la calidad y seguridad de los datos. Esto contiene la aplicación de criterios de calidad de datos determinados en ISO/IEC 25012:2008 (exactitud, completitud,

accesibilidad, consistencia) y el uso comprometido de la investigación académica bajo ISO/IEC 38505-1:2017.

De acuerdo con la naturaleza predictiva del problema, y gracias a que se podrá contar con la disponibilidad de grandes volúmenes de datos históricos, y con el propósito de comparar la eficacia entre diferentes modelos; esta investigación adoptará un enfoque cuantitativo, con un diseño no experimental de tipo predictivo y un método de análisis comparativo riguroso; el cual permitirá determinar las herramientas más factibles para predecir los estudiantes que se encuentran en riesgo de deserción.

2.9 HERRAMIENTAS

Esta selección de herramientas para el proyecto responde a la necesidad de comparar dos enfoques similares: el uso de modelos predictivos de minería de datos y estrategias de seguimiento académico tradicional, ambos métodos nos aportan varias ventajas, esto debido a que al ser constatados nos permiten obtener una visión más integral de los problemas en la deserción estudiantil. (Romero & Ventura,)

A continuación, se estarán listando las diversas herramientas que se agrupan en 4 categorías principales: **herramientas de minería de datos y machine *learning*** (ver tabla 1), **herramientas de seguimiento académico tradicional** (ver tabla 2), **estándares ISO** (tabla 3), metodología **Ágil Scrum** (ver tabla 4), **herramientas de planificación** (ver tabla 5) y herramientas básicas (ver tabla 6).

Tabla 1 - Herramientas de minería de datos y *machine learning*

#	Herramienta	Justificación de uso	Capacidades clave	Limitaciones
1	WEKA	Relevante para un entorno académico por su facilidad de uso y enfoque en estudiantes e investigadores con poca experiencia en programación.	Algoritmos de clasificación, *clustering* y validación cruzada integrados.	Escalabilidad limitada para grandes volúmenes de datos.

2	*RapidMiner*	Facilita la construcción de flujos de análisis sin necesidad de programar, lo que lo hace ideal para prototipos rápidos de modelos.	Comparación de modelos, integración con Python y R.	Versiones gratuitas restringen procesamiento avanzado.
3	Python (Scikit-learn, Pandas, *XGBoost*, etc.)	Herramienta flexible y estándar en proyectos de ciencia de datos, adecuada para análisis personalizados y escalables.	Algoritmos avanzados de ML, manipulación de datos, visualización.	Requiere conocimientos de programación.
4	R (con *RStudio*)	Aporta robustez estadística y librerías específicas para análisis educativo.	Modelos predictivos, validación cruzada, análisis multivariante.	Menos intuitivo para usuarios novatos.

Fuente: (Scherer et al., 2021)

Los sistemas de educación como SIGA permiten un control sistemático de calificaciones, asistencias y de inscripciones, generando datos confiables y bien estructurados, sin embargo su alcance es meramente descriptivo, solo muestran lo que ocurrió sin anticiparse a los riesgos (Arias Ortiz et al., 2021)

Tabla 2 - Herramientas de seguimiento académico tradicional

#	Herramienta	Justificación de uso	Capacidades clave	Limitaciones
1	SPSS	Muy usado en universidades, permite análisis estadísticos rápidos y confiables.	Regresión logística, análisis de cohortes y correlaciones.	Costo elevado y menor flexibilidad frente a Python o R.
2	STATA	Robusto en modelos longitudinales y análisis multivariados, útiles en seguimiento de cohortes estudiantiles.	Modelos de regresión avanzada, series de tiempo.	Licencia costosa y menos intuitiva que SPSS.
3	SIGA (Sistema de Gestión Académica, u otros similares)	Fuente directa de los datos institucionales (calificaciones, matrícula, asistencia).	Control de matrícula, historial académico.	Depende de la calidad de datos ingresados.

4	Encuestas y Entrevistas	Complementan los datos con factores socioeconómicos y motivacionales no capturados en sistemas formales.	Diagnóstico cualitativo.	Sesgo en respuestas y dificultad en escalabilidad.

Fuente: (*harleenk*, 2023)

Tabla 3 - Estándares ISO

#	Estándar	Relevancia para el proyecto	Alcance principal
1	ISO/IEC 25012:2008	Define dimensiones de calidad de datos esenciales para análisis predictivo confiable.	Exactitud, consistencia, accesibilidad.
2	ISO/IEC 20546:2019	Marco conceptual para *big data*, aplicable en minería de datos.	Terminología y guías para análisis.
3	ISO/IEC 38505-1:2017	Asegura gobierno responsable de datos en entornos educativos.	Lineamientos de gestión de información académica.
4	ISO/IEC 27001:2022	Estándar internacional de seguridad de la información.	Confidencialidad e integridad de datos estudiantiles.
5	ISO/IEC 27701:2019	Protección de datos personales y privacidad.	Extiende ISO 27001 hacia la gestión de datos sensibles.

Fuente: Elaboración propia

2.9.1 METODOLOGÍA ÁGIL SCRUM

El enfoque de la metodología Ágil Scrum aplicado a este proyecto de investigación nos permite adaptarnos a cambios en la disponibilidad de datos, la aplicación de nuevas técnicas de análisis y la validación de modelos predictivos, la ventaja de la implementación de Scrum es que promueve la colaboración constante, la retroalimentación continua y la entrega de resultados, que se adaptan al ciclo de desarrollo de los modelos de minería de datos y el análisis comparativo con el seguimiento académico tradicional (Hernández Cruz, 2022).

Tabla 4 - Metodología Ágil Scrum

#	Beneficios	Aplicación al proyecto

1	Flexibilidad	Ajustes en los modelos según disponibilidad de datos.
2	Retroalimentación continua	Validación iterativa de resultados parciales.
3	Reducción de riesgos	Identificación temprana de problemas metodológicos.
4	Resultados incrementales	Cada sprint aporta un avance en la comparación de modelos.

Fuente: (Adam Alami y Olivia Krancher, 2022)

Tabla 5 - Herramientas de planificación

#	Herramienta	Ventajas	desventajas	costo	Pertinencia en el proyecto
1	Jira	Muy usada en la industria	Curva de aprendizaje alta.	Gratis (máx. 10 usuarios). Plan Standard: $7.75/usuario/mes.	Ideal para gestión ágil del proyecto.
2	Trello	Colaboración sencilla	Funcionalidades limitadas en versión gratuita.	Gratis. Plan Premium: $10/usuario/mes.	Adecuado para gestión rápida de tareas.
3	Microsoft Project	Excelente para planificación formal	Poco flexible para metodologías ágiles,	$10–$55/usuario/mes (según plan).	Útil si se requiere reporte formal de hitos.
4	Miro	Excelente para lluvia de ideas	No es gestor de tareas completo.	Gratis. Plan Starter: $8/usuario/mes.	Útil para mapear causas de deserción (ej. Ishikawa).

Fuente: Elaboración propia

Tabla 6 - Herramientas de planificación

#	Herramientas	Funcionalidad	Justificación en el proyecto
1	Google Meet	Videoconferencias.	Reuniones de equipo y validación de resultados con expertos.
2	WhatsApp	Mensajería rápida.	Comunicación ágil en equipos pequeños.

Fuente: Elaboración propia

Tabla 7 - Herramientas básicas

#	Herramientas	Funcionalidad
1	Microsoft 365	mejorar la productividad con las aplicaciones innovadoras de Office, servicios inteligentes en la nube y seguridad de primer nivel.
2	Computadora	Ejecución de Software, creación de documentos y acceso a internet

Fuente: Elaboración propia

2.10 MARCO LEGAL

2.10.1 MARCO LEGAL NACIONAL

NORMATIVA LEGAL APLICABLE EN HONDURAS

Actualmente en nuestro territorio, el marco legal está orientado a proteger la igualdad en la educación y el acceso a oportunidades de formación. Según la Ley Fundamental de Educación la obligación del estado y las instituciones de velar por la permanencia y culminación de los estudios de los estudiantes, además el Consejo de Educación Superior (CES) garantiza la calidad educativa y el proceso de acompañamiento académico (UNAH, 2025a).

Tabla 8 - Normativa legal aplicable en Honduras

#	Ley	Año	Articulo	Descripción Articulo

1	Ley de Educación Superior, Decreto No. 121-1989	1989	Art. 3 y 4	Regula a las universidades privadas, garantizando su autonomía bajo la supervisión del CES, con obligación de asegurar calidad educativa y continuidad académica.
2	Estatutos del CES	2022	Varios	Define criterios de calidad, acreditación y procesos de acompañamiento académico aplicables a universidades privadas.

Fuente: Elaboración propia

NORMATIVA DE PROTECCIÓN DE DATOS PERSONALES

En la Ley de Protección de Datos Personales (Decreto No. 25-2017, 2017) establece que en Honduras el tratamiento de información sensible de los estudiantes, incluyendo registros académicos, socioeconómicos y personales se realice con el consentimiento informado y bajo los criterios de seguridad de la información. Esto implica que todas las instituciones que aplique a modelos predictivos deben proteger y resguardar la información para evitar vulneraciones de la privacidad estudiantil.

Tabla 9 - Normativa de protección de datos personales

#	Ley	Año	Descripción Articulo	Consentimiento informado requerido	Barreras e implicaciones para la minería de datos

1	Ley de Protección de Datos Personales	25-2017	La presente Ley es de orden público y de observancia general en toda la República y tiene por objeto la protección de los datos personales, garantizando su confidencialidad, integridad y acceso adecuado	1. La ley demanda que todo tratamiento de datos personales sensibles (como historial académico, variables socioeconómicas o indicadores de desempeño) se realice con el consentimiento previo, libre y explícito del estudiante. 2. El consentimiento debe detallar el propósito del análisis predictivo, el alcance de la investigación, la forma en que se almacenarán los datos y los mecanismos de resguardo.	1. Limitación en la cantidad de datos: si un estudiante no otorga su consentimiento, sus registros no pueden incluirse en el modelo, lo que puede afectar la representatividad del *dataset*. 2. Costos de cumplimiento: las universidades deben implementar políticas de seguridad de la información, cifrado y protocolos de acceso restringido, lo cual aumenta los recursos necesarios para la investigación.

Fuente: Elaboración propia

MARCO REGULATORIO PARA LA EDUCACIÓN

En el marco legal para la educación superior en Honduras se encuentra regulado por la Ley de Educación Superior (Decreto No. 121-1989, 1989), y sus cambios posteriores, la cual nos establece la autonomía universitaria, también la obligación de garantizar la calidad educativa y la inclusión, además la Dirección de Educación Superior(DES) de la Secretaria de Educación supervisa las políticas relacionadas con la continuidad estudiantil, asegurando que las instituciones cuentan con mecanismos de tutoría, orientación y acompañamiento académico para evadir la deserción estudiantil.

Tabla 10 - Normativa de educación superior

#	Ley	Año	Descripción Articulo
1	Ley de educación superior	121-1989	establece el marco jurídico del nivel universitario en Honduras, otorgando a la UNAH la función exclusiva de dirigir y desarrollar la educación superior.

Fuente: Elaboración propia

2.10.2 MARCO LEGA INTERNACIONAL

2.4.1 Consideraciones éticas en la aplicación de modelos predictivos

En el uso de modelos en la minería de datos en la educación superior tiene un compromiso ético con la seguridad de la información, transparencia e igualdad en la interpretación de los

resultados, tenemos que evadir sesos en los algoritmos que puedan generar cierta discriminación hacia estudiantes por razones socioeconómicas, procedencia geográfica o género, La (UNESCO, 2025) nos recomienda que las universidades adopten los principios de ética digital, para garantizar que los modelos predictivos utilicen apoyo para evitar que estos se conviertan en mecanismos de exclusión académica.

Tabla 11 - Normativas internacionales

#	Ley	Año	Descripción	Artículo
1	Declaración Universal de los Derechos Humanos (ONU)	1948	Reconoce la educación como derecho humano fundamental.	Art. 26: La educación superior debe ser accesible a todos, en función de los méritos.
2	Pacto Internacional de Derechos Económicos, Sociales y Culturales (PIDESC)	1966	Compromete a los Estados a garantizar el derecho a la educación, fomentando la accesibilidad progresiva y gratuita de la educación superior.	Art. 13: Derecho a la educación, igualdad de acceso a la educación superior.
3	Declaración Mundial sobre la Educación Superior en el Siglo XXI (UNESCO)	1998	Establece principios de calidad, pertinencia y equidad en la educación superior, instando a políticas contra la deserción.	Art. 1 y 2: La educación superior como derecho y bien público.
4	Agenda 2030 – Objetivos de Desarrollo Sostenible (ONU)	2015	Plantea metas globales para garantizar educación inclusiva, equitativa y de calidad.	ODS 4: Meta 4.3 – acceso igualitario a la educación superior asequible y de calidad.
5	Declaración de la Conferencia Regional de Educación Superior (CRES, UNESCO-IESALC)	2018	Reconoce la educación superior como bien público y derecho humano; resalta la obligación de reducir la exclusión y la deserción.	Principios Generales: Equidad, inclusión y permanencia en la educación superior

Fuente: Elaboración propia

CAPÍTULO III. METODOLOGÍA

3.1 CONGRUENCIA METODOLÓGICA

La congruencia metodológica es crucial ya que constituye el elemento que rige el proceso de la investigación. Asegurando que el título de la investigación, el problema de investigación, los objetivos, la metodología, instrumentos y variables mantenga una coherencia y lógica con base en el análisis predictivo. A continuación, en la matriz metodológica se aprecia en detalle cada uno de los componentes descritos anteriormente.

3.1.1 MATRIZ METODOLÓGICA

Título de investigación	Preguntas de investigación	Objetivos de la investigación		Variables	Indicadores
		General	Específicos		
Análisis predictivo del riesgo de deserción en estudiantes de primer ingreso de la Universidad Tecnológica Centroamericana	En estudiantes de primer ingreso (P), ¿en qué medida los factores más predictivos del riesgo de deserción (O), al ser analizados mediante modelos predictivos (árboles de decisión y regresión logística) (I), permiten una identificación más temprana y precisa del riesgo de deserción (C)?	Evaluar la efectividad de técnicas de minería de datos y *learning analytics* para predecir y anticipar el riesgo de deserción en estudiantes de primer ingreso, en comparación con el seguimiento académico tradicional.	Identificar (S) los factores más predictivos del riesgo de deserción en estudiantes de primer ingreso (M) mediante la aplicación de modelos predictivos como árboles de decisión y regresión logística, (A) utilizando datos históricos disponibles de la Universidad Tecnológica Centroamericana, (R) para una determinación más temprana y precisa de este, y (T) al finalizar el análisis de dichos datos históricos.	Factores académicos, socioeconómicos y de rendimiento que influyen en la deserción.	Porcentaje de estudiantes que no se matriculan en el siguiente periodo Tiempo transcurrido hasta el abandono
	En estudiantes de primer ingreso (P), ¿qué precisión		Determinar la precisión predictiva y la capacidad de	Precisión predictiva y capacidad de	Porcentaje de aciertos de predicción

	predictiva y capacidad de anticipación demuestran los modelos predictivos (árboles de decisión y regresión logística) (I) para identificar el riesgo de deserción (O), en contraste con la capacidad de detección del seguimiento académico tradicional (C)?		anticipación de los modelos de árboles de decisión y regresión logística para identificar el riesgo de deserción.	anticipación de los modelos.	Anticipación en semanas
	En estudiantes de primer ingreso (P), ¿de qué manera la información generada por un modelo predictivo (árboles de decisión y regresión logística) (I) puede optimizar el diseño y la pertinencia de las estrategias de acompañamiento estudiantil para reducir el riesgo de deserción (O), en comparación con la información obtenida del seguimiento académico tradicional (C)?		Analizar cómo la información generada por los modelos predictivos puede optimizar el diseño y la pertinencia de las estrategias de acompañamiento estudiantil.	Pertinencia y optimización de estrategias de acompañamiento.	Número de iteraciones Tiempo de procesamiento Número de intervenciones realizadas Tasa de retención

3.1.2 ESQUEMA DE VARIABLES DE ESTUDIO

Tipo	Variable
Variable Dependiente	Deserción estudiantil
Variables Independientes (Predictoras)	Factores académicos
	Factores socioeconómicos
	Factores de rendimiento
Métricas de Evaluación	Precisión predictiva y capacidad de anticipación
	Optimización del diseño
	Pertinencia de estrategias de acompañamiento estudiantil

La variable dependiente corresponde a la deserción estudiantil, definida como el abandono definitivo o temporal de los estudios por parte de los estudiantes y corresponde al elemento central de nuestro análisis mediante la determinación de los patrones o causas de este abandono.

En relación con las variables independientes que son las predictoras corresponden a los factores que inciden o están asociados a la deserción. Por su parte, las métricas como; precisión predictiva y capacidad de anticipación, optimización del diseño y pertinencia de estrategias de acompañamiento estudiantil; son fundamentales para evaluar la eficacia del modelo de predicción y la utilidad de las variables seleccionadas.

3.1.3 OPERACIONALIZACIÓN DE LAS VARIABLES

Variable	Definición conceptual	Definición operacional	Dimensiones	Indicadores	Ítems
Deserción estudiantil	Es el abandono, temporal o definitivo, de un programa académico antes de su conclusión, causado por factores académicos, sociales e institucionales (Miño de Gauto, 2021).	Registro institucional de baja académica durante el primer año.	Continuidad de matrícula Permanencia académica	Porcentaje de estudiantes que no se matriculan en el siguiente periodo Tiempo transcurrido hasta el abandono	¿Ha abandonado temporal o definitivamente sus estudios? ¿Motivo principal del abandono?
Precisión predictiva y capacidad de anticipación	La precisión predictiva se refiere al grado en que un modelo analítico anticipa	Cálculo de métricas de rendimiento del modelo y tiempo promedio de	Exactitud del modelo Tiempo de anticipación	Porcentaje de aciertos de predicción Anticipación en semanas	¿Con cuánta anticipación se detectó el riesgo?

	correctamente la deserción; mientras que la capacidad de anticipación es la detección temprana del riesgo (Zárate-Valderrama et al., 2021).	anticipación antes del abandono.			
Optimización del diseño	Proceso de ajuste y mejora sistemática del modelo predictivo para maximizar su desempeño en la identificación de estudiantes en riesgo (Bellaj et al., 2024)	Ajuste iterativo de parámetros y selección de variables relevantes dentro del proceso de modelado.	Ajuste de parámetros Rendimiento del modelo	Reducción del error de predicción tras N ciclos de reentrenamiento Tiempo de procesamiento	¿Huno mejora en la precisión después de la optimización?
Pertinencia de estrategias de acompañamiento estudiantil	Grado en que los programas de apoyo académico se adecuan al perfil, necesidades y factores de riesgo identificados en los estudiantes en riesgo de deserción (Marrón Ramos et al., 2022)	Implementación de acciones diferenciadas para casos identificados como vulnerables según el modelo predictivo.	Relevancia de las acciones de apoyo Eficacia en la retención	Número de intervenciones realizadas Tasa de retención	¿Ayudó la medida a su permanencia?

Fuente: Elaboración propia

3.1.4 HIPÓTESIS

H1-Hipótesis alternativa = Los modelos predictivos basados en las técnicas de minería de datos (regresión logística y árboles de decisión) alcanzan una capacidad de predicción del riesgo de deserción estudiantil en estudiantes de primer ingreso, que es igual o superior al 85%, superando la precisión de los métodos convencionales de seguimiento académico.

HO-Hipótesis nula = No existe diferencia estadísticamente significativa en la precisión para identificar el riesgo de deserción estudiantil entre los modelos predictivos basados en minería de datos (regresión logística y arboles de decisión) y los métodos convencionales de seguimiento académico para identificar el riesgo de deserción estudiantil.

Una investigación sobre el diseño e implementación de una red neuronal artificial (RNA) para predecir el rendimiento académico de los estudiantes de Ingeniería Civil específicamente del curso de Física de la Universidad Nacional Intercultural Fabiola Salazar Leguía, ubicada en Bagua-Perú, empleando datos históricos. Determinó mediante una comparación entre dos algoritmos de entrenamiento, el Levenberg-Marquardt, que mostró un 86% de capacidad predictiva, y el *Scaled Conjugated Gradient*, con un 70%. Los hallazgos indican que es posible lograr una efectividad del 88.67%, rebasando de manera notable el umbral del 85% sugerido en su hipótesis. El estudio concluye, en última instancia, que las RNA son un instrumento potente para hacer predicciones en el ámbito educativo. (Incio Flores et al., 2021)

3.2 ENFOQUE O TIPO DE INVESTIGACIÓN

La investigación tiene un enfoque cuantitativo, basado en el positivismo desde una perspectiva epistemológica; esta corriente defiende que el conocimiento acerca de los fenómenos sociales es posible de manera objetiva a través del estudio sistemático y la observación de datos numéricos. La ontología que fundamenta esta perspectiva se basa en la suposición de que hay realidades sociales que pueden ser observadas y medidas y que no están sometidas a la percepción individual; esto es particularmente relevante para el análisis del riesgo de deserción estudiantil.

La epistemología cuantitativa da preferencia a la objetividad, la verificabilidad y la posibilidad de replicar los resultados, lo cual facilita el análisis de patrones y conexiones a través de modelos predictivos y métodos estadísticos. Este método es apropiado, ya que la pregunta principal de investigación consiste en identificar y examinar los factores relacionados con el rendimiento académico mediante la recolección de datos objetivos (como calificaciones, historial académico, métricas de riesgo, etc.) para comprobar hipótesis acerca del abandono estudiantil a partir de patrones detectables. En la literatura metodológica, muchos autores aconsejan utilizar el enfoque cuantitativo para analizar fenómenos como el abandono de los estudiantes, sobre todo cuando se utilizan modelos predictivos que pueden ser medidos con métricas claras, comparables y replicables. Romero & Ventura (2020) señalan que “el análisis cuantitativo permite no solo la observación de correlaciones, sino también la previsión de situaciones”.

3.3 ALCANCE

La presente investigación tiene un alcance predictivo, comparativo y correlacional. Ya que el propósito principal es construir un modelo que anticipe el riesgo de deserción estudiantil a partir

de datos históricos, además, de determinar las relaciones entre diferentes variables (como el índice académico, el número de clases reprobadas, asignaturas cursadas, calificaciones obtenidas, asistencia, clases retiradas, entre otras).

El alcance correlacional permite establecer la dirección e intensidad de las relaciones entre variables significativas, mientras que el predictivo utiliza esas relaciones para anticipar los posibles resultados futuros. Autores como Hernández-Sampieri & Mendoza, (2020) y V. Flores et al., (2022) "destacan que el alcance predictivo y correlacional es más pertinente en investigaciones donde existen antecedentes cuantitativos previos y se dispone de suficiente información para construir y validar modelos predictivos".

Al realizar la investigación en la Universidad Tecnológica Centroamericana, y emplear técnicas avanzadas como minería de datos y análisis de aprendizaje, se logra una aproximación profunda y específica, que no solo describe el problema, sino que también proporcione información útil para la toma de decisiones institucionales. La literatura acerca de la deserción estudiantil respalda la transición entre enfoques descriptivos y exploratorios, hacia investigaciones predictivas que permitan pronosticar de forma oportuna el riesgo de deserción estudiantil.

3.4 DISEÑO

En coherencia con el alcance, se empleó un diseño de investigación no experimental de tipo predictivo, ya que se trabajó con datos históricos de los últimos dos años, tanto para las calificaciones y el año de ingreso de los estudiantes, además, se basó en los datos tal cual se presentan sin ninguna manipulación de las variables, con el propósito de validar modelos de minería de datos que permitan anticipar el riesgo de deserción en estudiantes de primer ingreso de la Universidad Tecnológica Centroamericana.

Por lo cual, la investigación se realizó bajo un enfoque cuantitativo, utilizando técnicas de análisis predictivo como regresión y árbol de decisión para identificar patrones y relaciones entre las bases de datos de la institución.

3.4.1 POBLACIÓN

La población estudiada está constituida por los estudiantes de primer ingreso de la Universidad Tecnológica Centroamericana; entre los años 2025 y 2023.

Tabla 12 - Población estudiantes matriculados

Año	Estudiantes matriculados de primer ingreso
2025 – Q3	1267
2025 – Q2	2651
2025 – Q1	1469
2024 – Q4	2208
2024 – Q3	1440
2024 – Q2	2787
2024 – Q1	1382
2023 – Q4	2293
Total	15497

Fuente: Elaboración propia

Según los registros institucionales, este grupo asciende a N = 15,497 estudiantes, esta delimitación responde al objetivo central de la investigación que es el de analizar el riesgo de deserción en estudiantes de primer ingreso.

Justificación de la delimitación

La relevancia académica para los estudiantes de primer ingreso, es que corresponde a un período crítico en la trayectoria estudiantil, ya que están concentrados los mayores índices de deserción y reprobación, por lo que constituye el punto clave para la detección temprana del riesgo de deserción (Moya, E., 2021).

Para la variable predictora clave se tomó en cuenta la reprobación de al menos una asignatura, una clase retirada, una clase en las que los estudiantes quedaron sin derecho o la no matrícula en un periodo consecutivo, las cuáles se identificaron como un predictores significativos de abandono, lo que nos ayuda a reforzar su selección como criterio de inclusión (González-Nucamendi et al., 2023).

Para la disponibilidad y confiabilidad de datos se contó con acceso completo a los registros académicos institucionales, lo que nos permitió trabajar con la totalidad de la población sin necesidad de muestreo, este enfoque tipo censo garantiza la representatividad de la información y la precisión de los modelos de minería de datos y análisis predictivo. Por lo cual, se hizo uso de toda la población; aplicando el censo de datos, ya que es ideal para modelos predictivos.

Los criterios de inclusión se definieron de la siguiente manera:

1. Población objetivo: Estudiantes de primer ingreso matriculados en 8 periodos de los años 2023, 2024 y 2025.

2. Condición académica: Son estudiantes de los cuales se registra al menos una clase reprobada, una clase retirada, una clase en la que perdieron derecho, en los periodos de análisis, ya que se identificaron estas variables como predictoras relevante ante un riesgo académico y deserción (González-Nucamendi et al., 2023)
3. Disponibilidad de información: Fueron considerados los estudiantes con datos académicos completos y registrados en la base de datos del sistema institucional, incluyendo historial de notas, asignaturas cursadas y calificaciones obtenidas, para reducir los márgenes de error en los cálculos.

Se establecieron los criterios de exclusión de información:

- Los estudiantes de reingreso, equivalencias o transferencias externas, ya que sus trayectorias no corresponden al perfil de ingreso directo de primer año y puede distorsionar la información (Yu et al., 2021).
- Estudiantes con registros incompletos o inconsistentes en las bases de datos, esto podría sesgar los resultados del análisis predictivo.

En este sentido, la población estudiada está conformada por todos los estudiantes de primer ingreso, mediante el análisis de 8 periodos académicos comprendidos entre los años 2023 al 2025. Incluyendo únicamente a los estudiantes con registros completos en las bases de datos, y excluyendo los registros incompletos o que no cumplan con la condición de primer ingreso.

Utilizando el censo, ya que se incluyó el total de la población para obtener datos más fiables debido a las herramientas de minería de datos que se emplearon.

La decisión de emplear el censo es porque permite una mayor precisión en la predicción y gracias a la disponibilidad que se tiene a los datos de los estudiantes y para contar con un conjunto más amplio de datos lo que permitió un mejor análisis predictivo. Garantizando de esta forma una representatividad de la información y por ende la confiabilidad de los resultados.

3.4.2 MUESTRA

Dado que la investigación tiene un enfoque predictivo y para garantizar la representatividad de los resultados y maximizar el rendimiento predictivo, se hará uso de toda la población. Como indica Kuhn & Johnson, (2013) “con tamaños de muestras pequeños, los valores que aparentan ser

atípicos pueden ser el resultado de una distribución sesgada, ya que no se dispone de datos suficientes, mientras que un análisis predictivo con datos completos mejora la precisión y generalización". Ya que, se contó con el acceso completo a las bases de datos de los estudiantes de primer ingreso de 8 periodos académicos comprendidos entre los años del 2023 al 2025, lo permitió que todos los estudiantes fueran incluidos mejorando la modelación en la detección de patrones reales de deserción estudiantil, asegurando la eliminación del error muestral.

3.4.3 TÉCNICAS, INSTRUMENTOS Y PROCEDIMIENTOS APLICADOS

Para la investigación se utilizó el enfoque cuantitativo de tipo no experimental, transversal y correlacional, que está sustentado en las técnicas de minería de datos y análisis predictivo para identificar factores asociados al riesgo de deserción estudiantil, la técnica principal es el análisis estadístico de aprendizaje automático, específicamente utilizando la regresión logística y árboles de decisión, por su capacidad de modelar relaciones entre variables categóricas y continuas (Cristobal Romero & Ventura, 2020).

El instrumento de recolección fue la base de datos académica institucional, la cual contiene registros históricos de estudiantes de primer ingreso de entre el 2023 y el 2024, esta base incluye variables académicas. Antes de realizar el análisis los datos fueron anonimizados y se aplicó mediante un proceso de limpieza, validación de integridad, tratamiento de valores perdidos y codificación de variables, al seguir recomendaciones de (Cristobal Romero & Ventura, 2020) para la minería de datos educativa.

A continuación, se muestran las fases secuenciales con las que se inició el procedimiento:

- Extracción y depuración de datos: mediante el sistema institucional, con la autorización académica.
- División de la base en conjuntos de entrenamiento (70%) y prueba (30%) para la evaluación posterior de la capacidad de generalización de los modelos.
- Entrenamiento de modelos predictivos (regresión logística y árboles de decisiones) mediante software de análisis estadístico y aprendizaje automático, como R o Python.

- Evaluación de desempeño utilizando métricas de precisión, sensibilidad, especificidad, área bajo curva ROC y capacidad de anticipación temporal (Quintana, Llatasi, 2024).

3.5 FUENTES DE INFORMACIÓN

3.5.1 FUENTES PRIMARIAS

Debido a que el estudio es un análisis predictivo sobre la deserción estudiantil y dado que la principal fuente de información corresponde a las bases de datos académicas institucionales de la Universidad Tecnológica Centroamericana las que constituyen una fuente de información secundaria. Además, siendo el principal propósito de la investigación modelar y predecir el comportamiento mediante el análisis de registros históricos reales, y no sobre la recolección de percepciones; no se emplearon las fuentes primarias ya que estas no se integran directamente al entrenamiento del algoritmo por su naturaleza subjetiva y pueden comprometer la estabilidad y replicabilidad del modelo.

3.5.2 FUENTES SECUNDARIAS

La investigación se fundamentó en las bases de datos académicas institucionales de la Universidad Tecnológica Centroamericana objeto de estudio, las cuales constituyen una fuente de información secundaria y la principal de evidencia empírica, ya que estas concentran los registros históricos de los estudiantes de primer ingreso, entre ellos, el historial académico, número de asignaturas reprobadas, promedio global, carga académica, clases retirados y clases sin derecho. Tal como menciona Romero & Ventura, (2020b) “la construcción de modelos predictivos prioriza el uso de datos históricos objetivos y estructurales, sin aplicación de instrumentos de recolección primaria, cuyo carácter es subjetivo y transversal pueden limitar la estabilidad del modelo”.

La mayoría de los estudios predictivos hacen uso de datos académicos institucionales, y el valor de origen metodológico de esta fuente radica en que nos proporciona datos objetivos, verificables y completos, que terminan siendo una condición indispensable para la minería de datos y el análisis predictivo. Solo mediante la información oficial, capturada en el momento del ingreso y durante el primer año ingreso, lo cual nos permitió garantizar cierta validez interna y replicabilidad de los modelos de regresión logística y árboles de decisiones que se aplicaron para identificar factores de riesgos de deserción.

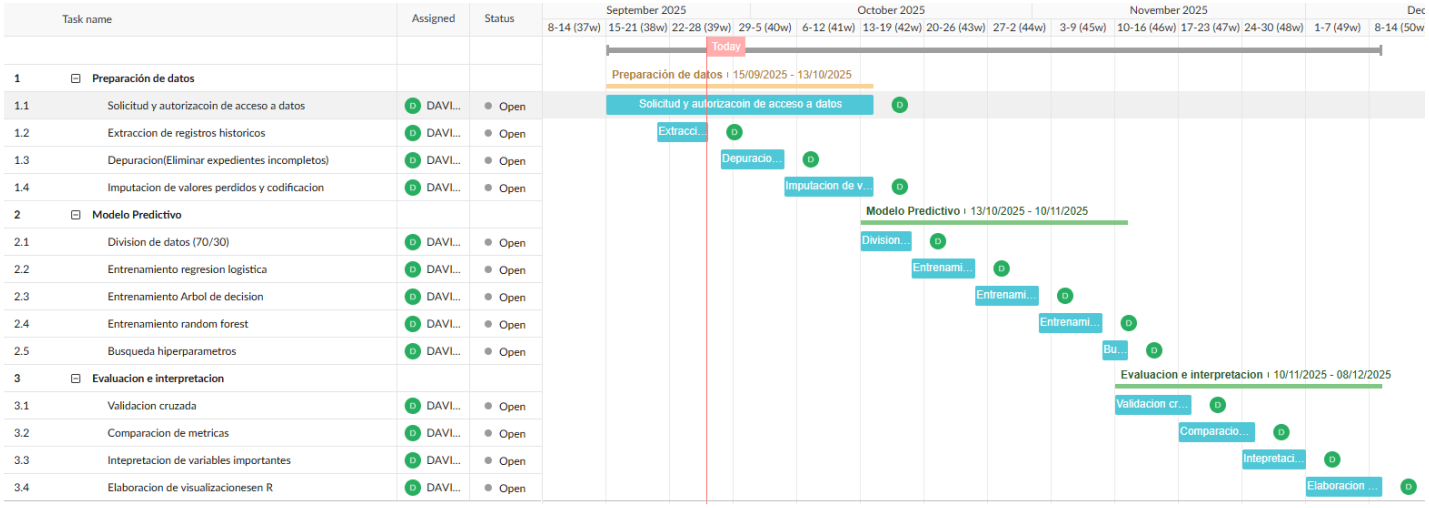
3.6 PLAN DE ANÁLISIS

Fase	Actividad	Descripción	Entregable
Preparación de datos	Recolección y autorización de datos	Gestión de permisos institucionales y extracción de registros históricos de estudiantes de primer ingreso.	Base de datos institucional autorizada y exportada.
	Depuración y anonimización	Eliminación de expedientes incompletos o inconsistentes, tratamiento de valores perdidos mediante imputación estadística y anonimización para cumplir normativa de protección de datos (INE, 2023a).	*Dataset* limpio, anonimizado y listo para análisis.
	Codificación y normalización	Transformación de variables categóricas a formato numérico y estandarización de escalas para análisis predictivo.	*Dataset* codificado y normalizado.
Modelado predictivo	Validación cruzada	Aplicación de validación cruzada estratificada de k pliegues (k = 10) sobre el conjunto de datos, con el objetivo de evaluar el desempeño de los modelos de manera robusta y reducir el sesgo asociado a una única partición de los datos.	Esquema de validación cruzada definido.
	Entrenamiento de modelos	Implementación de algoritmos supervisados: regresión logística, árboles de decisión y *random forest* siguiendo recomendaciones de (Bellaj et al., 2024; C. Romero & Ventura, 2020).	Modelos entrenados (tres variantes).
	Ajuste de hiperparámetros	Optimización de los modelos con búsqueda en malla (*grid search*) y validación cruzada de 10 pliegues para reducir sobreajuste, según (Sohil et al., 2022)	Modelos optimizados y listos para evaluación.
Evaluación e interpretación	Evaluación de desempeño	Cálculo de métricas estándar (exactitud, precisión, exhaustividad, F1 y AUC-ROC) y comparación de resultados entre algoritmos.	Tabla de métricas comparativas.

	Interpretación de variables	Cálculo de la importancia de cada predictor y análisis de factores que más contribuyen al riesgo de deserción.	Ranking de variables predictoras clave.
	Visualización y reporte	Elaboración de gráficos en R y redacción del informe final de hallazgos y recomendaciones para la universidad.	Informe de resultados y visualizaciones gráficas.

Fuente: Elaboración propia

Ilustración 2 - Planeación de la investigación



Fuente: Elaboración propia

CAPÍTULO IV. RESULTADOS Y ANÁLISIS

En este capítulo se exponen los resultados obtenidos durante el proceso de investigación a través del análisis de los datos recolectados mediante la aplicación de técnicas de minería de datos y modelos predictivos; con el propósito de dar respuesta a la pregunta de investigación sobre la predicción de la deserción estudiantil y verificar el cumplimiento de los objetivos propuestos, este capítulo, presenta el análisis empírico de los datos académicos.

Para lograr este fin, el capítulo se estructura en dos pilares fundamentales; el primero, presenta un Análisis Exploratorio de Datos (AED) (sección 4.1) que muestra los patrones y características inherentes a la data; el cual se basa en un enfoque cuantitativo, utilizando datos objetivos y verificables obtenidos de la base de datos de estudiantes. Posteriormente en el segundo pilar y para ofrecer un contexto completo sobre la procedencia y el rigor metodológico de la información analizada, se detalla el Informe del Proceso de Recolección de Datos (sección 4.2), asegurando la transparencia y replicabilidad del estudio.

La exposición culmina con los Resultados y Análisis de las Técnicas Aplicadas (sección 4.3). En esta sección se aborda la evaluación de la efectividad de la aplicación de técnicas de minería de datos y *learning analytics* para predecir y anticipar el riesgo de deserción en estudiantes.

De esta manera, la adopción del enfoque cuantitativo y el diseño no experimental de tipo predictivo se materializa a través de los datos analizados en el AED y la posterior evaluación de los modelos predictivos propuestos (árboles de decisión y regresión logística), con el fin de fortalecer las estrategias de apoyo académico.

4.1 ANÁLISIS EXPLORATORIO DE DATOS (AED)

La investigación se basa en un enfoque cuantitativos y utiliza datos objetivos y verificables proveniente de una base de datos proporcionada por la universidad. El conjunto de datos con el que se trabajará es de 15,497 estudiantes.

Entre las variables analizadas tenemos; la deserción estudiantil, los factores académicos, así como los socioeconómicos, y de rendimiento. Además, otra de las variables es la precisión predictiva y capacidad de anticipación; también, la optimación del diseño y la pertinencia de estrategias de acompañamiento estudiantil.

A continuación, se muestra la clasificación de las variables entre cuantitativas y cualitativas, la cual, es primordial para definir los métodos de análisis más adecuados, para el análisis estadístico mediante las variables cualitativas; mientras que las variables cualitativas permiten el análisis profundo de las percepciones, contextos y condiciones implícitos del estudio.

Variables Cuantitativas	Variables Cualitativas
Deserción estudiantil	Factores académicos
Factores de rendimiento	Factores socioeconómicos
Precisión predictiva y capacidad de anticipación	Optimización del diseño
	Pertinencia de estrategias de acompañamiento estudiantil

Fuente: Elaboración propia

4.1.1 DESCRIPCIÓN GENERAL DEL CONJUNTO DE DATOS

Esta sección presenta las características fundamentales del conjunto de datos, que constituye la base empírica para el desarrollo de los modelos predictivos relacionados con la deserción. La investigación se centra en un método cuantitativo, utilizando datos verificables y objetivos que se extraen de las bases de datos académicas de la universidad.

Puesto que se cuenta con el acceso completo de los registros y basado en la naturaleza del análisis predictivo, se optó por un análisis de la población total accesible, para maximizar la potencia estadística, donde la cifra total de registros que se emplean en el Análisis Exploratorio de Datos (AED) asciende a 15,497. Los datos de este estudio se extrajeron del sistema de información estudiantil de la Universidad Tecnológica Centroamericana, que se examina.

Tabla 13 - Características de la población de estudio

Categoría	Descripción
Números de registros	Es de 15,497 registros
Número de variables	El análisis se basó en 7 variables, 3 cuantitativas y 4 cualitativas.
Número de periodos	Se evaluaron 8 periodos académicos

Fuente: Elaboración propia

4.1.1.1 ESTRATEGIA DE INVESTIGACIÓN: EL CENSO

Para el presente estudio, se analizará toda la población, sin aplicar técnicas de muestreo, ya que la cantidad del grupo de estudiantes es controlable, es por ello, que se ha adoptado una estrategia de censo para trabajar con la totalidad de la población definida (N = 15,497 estudiantes).

En el campo de la estadística, una investigación se define como censo cuando esta es exhaustiva y trabaja con todas las unidades poblacionales, a diferencia del muestreo. Se optó por el censo, dadas las facilidades de acceso completo a los registros institucionales, con el objetivo principal de maximizar la potencia predictiva y estadística.

La justificación más sólida para la realización de un censo radica precisamente en que elimina el error de muestreo. Cabe mencionar, que el muestreo, al ser una investigación parcial, busca inferir las propiedades de una población a partir de una porción de ella, y esto inherentemente conlleva un error de muestreo que, si bien es medible y acotable como señala (Gutiérrez Rojas, 2016), existe.

En contraste, el censo, al recoger datos de todos los elementos de la población y abarcar la totalidad del campo de observación (Martínez Bencardino, 2019), no necesita realizar inferencias a partir de una muestra y dado que la investigación tiene el propósito de identificar patrones específicos de irregularidades, incluyendo datos atípicos trabajar con toda la población permite que no se pierda información relevante. Esto significa que garantiza la obtención del valor exacto del parámetro poblacional, sin la presencia del error de muestreo, lo que lo convierte en un método de recolección de datos extremadamente preciso, este enfoque resulta particularmente adecuado, ya que el objetivo no es generalizar los datos, sino documentar de forma precisa las irregularidades presentes en contextos y periodos determinados.

No obstante, es importante reconocer que los hallazgos están limitados a la población analizada y no pueden generalizarse automáticamente a otros contextos o periodos. Con el fin de minimizar posibles sesgos, se implementará una revisión exhaustiva de los registros cruzando datos con múltiples fuentes internas asegurando que no existen omisiones significativas. De esta manera, aunque no se utilice muestreo, el diseño garantiza validez interna para los fines específicos del estudio, cumpliendo con los estándares metodológicos establecidos para investigaciones con poblaciones delimitadas.

Categorizar las variables es un paso esencial del Análisis Exploratorio de Datos (AED) y establece la estrategia para limpiar y preparar los datos. Según las observaciones, estos procesos deben ser elaborados y documentados con precisión en términos metodológicos.

La siguiente clasificación de las variables ofrecidas se basa en sus características cualitativas y cuantitativas:

Variables numéricas (cuantitativas)

Estos factores, que son las variables fundamentales para el modelado predictivo, pueden ser analizados mediante la estadística descriptiva.

Variable	Tipo de Escala	Justificación y Relación
índice_general	Continua	Mide el rendimiento académico principal. Es análogo a la variable
índice_graduación	Continua	Mide el progreso del rendimiento acumulado para la graduación
edad	Discreta	Factor demográfico. Esta variable es crítica para el manejo de valores atípicos
clases_reprobadas	Discreta	Conteo directo de riesgo académico. Esta variable es señalada explícitamente en la tesis como discreta y es un factor predictivo central del estudio.
clases_matriculadas	Discreta	Número de clases que el estudiante tomó en el periodo.
clases_sin_derecho	Discreta	Conteo de asignaturas perdidas por inasistencia u otra razón disciplinaria.
clases_retiradas	Discreta	Conteo de asignaturas retiradas formalmente por el estudiante.

Fuente: Elaboración propia

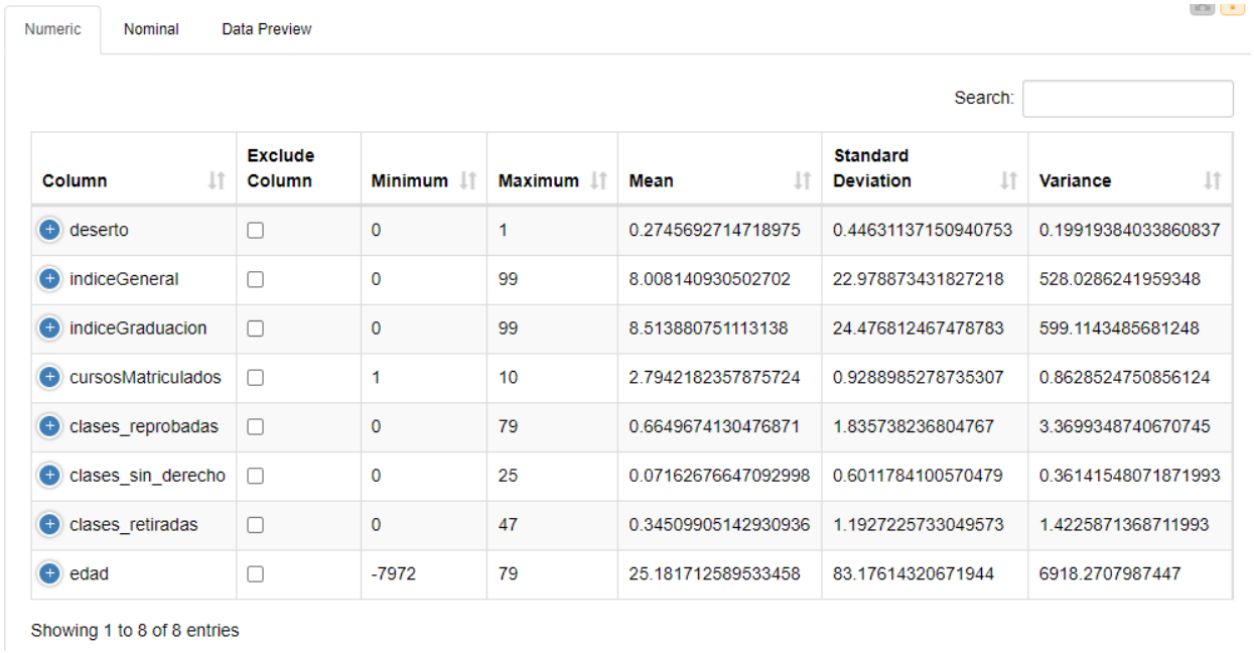
Variables Cualitativas (Categóricas)

Estas variables sirven para dividir y clasificar a los alumnos. Su esencia requiere procesos de transformación de datos antes de que sean integradas en los modelos de regresión.

Variable	Tipo de Escala	Clasificación y Función
desertor_temprano	Binaria (Objetivo)	Es la variable dependiente central que el estudio busca predecir. Es análoga a la variable 'deserción', clasificada como binaria. Se define operacionalmente como el abandono definitivo o temporal de los estudios.
nombre_carrera	Nominal	Clasifica el programa de estudio. Es análoga a la variable 'carrera', una variable categórica que debe transformarse mediante codificación *one-hot-encoding* para los modelos de regresión.
facultad	Nominal	Clasifica la división académica del estudiante.
campus_original	Nominal	Clasifica por la sede de la universidad.
sexo	Nominal	Clasifica por género del estudiante.
nacionalidad	Nominal	Factor demográfico.
tipo_ingreso	Nominal	Clasificación primaria del modo de entrada a la universidad.
código_carrera	Nominal	Identificador de la carrera que no posee valor matemático, solo clasificatorio.
periodo	Nominal/Ordinal	Identificador del tiempo académico cursado.
nivel	Nominal/Ordinal	Nivel de estudio (ej. pregrado).
cuenta	Nominal (Identificador)	Es el identificador único del estudiante; es tratado como cualitativo, aunque sea numérico, ya que no se realizan operaciones matemáticas con él.

Fuente: Elaboración propia

Sintetiza las medidas de dispersión (mínimo, máximo y desviación estándar) y las medidas de tendencia central (mediana y media) para las variables cuantitativas que se emplearon en el Análisis Exploratorio de Datos (AED).



Numeric | Nominal | Data Preview

Search:

Column	Exclude Column	Minimum	Maximum	Mean	Standard Deviation	Variance
deserto	☐	0	1	0.2745692714718975	0.44631137150940753	0.19919384033860837
indiceGeneral	☐	0	99	8.008140930502702	22.978873431827218	528.0286241959348
indiceGraduacion	☐	0	99	8.513880751113138	24.476812467478783	599.1143485681248
cursosMatriculados	☐	1	10	2.7942182357875724	0.9288985278735307	0.8628524750856124
clases_reprobadas	☐	0	79	0.6649674130476871	1.835738236804767	3.3699348740670745
clases_sin_derecho	☐	0	25	0.07162676647092998	0.6011784100570479	0.36141548071871993
clases_retiradas	☐	0	47	0.34509905142930936	1.1927225733049573	1.4225871368711993
edad	☐	-7972	79	25.181712589533458	83.17614320671944	6918.2707987447

Showing 1 to 8 of 8 entries

Ilustración 3 - Vista preliminar de métricas estadísticas de variables cuantitativas

El análisis preliminar reveló valores consistentes en la mayoría de las variables; no obstante, la variable edad mostró un valor mínimo de -7,972, lo que indica un error en la captura de datos que distorsionó su distribución y subraya la importancia de realizar una depuración previa antes de proceder con análisis posteriores.

4.1.2 LIMPIEZA Y PREPARACIÓN DE LOS DATOS

Con el fin de asegurar la integridad y calidad de los registros de los estudiantes de primer año, se realizó el proceso de limpieza y preparación de datos a través de un flujo sistemático y paso a paso utilizando la herramienta KNIME y usando librerías de Python. El propósito principal fue garantizar que los datos empleados en los modelos de árboles de decisión y regresión logística fueran válidos.

Las columnas categóricas no aplican para valor atípico (*outliers*) son: campus_original, periodo, desertor_temprano, sexo, nivel, tipo_ingreso_1, tipo_ingreso_2, código_carrera, nombre_carrera, facultad.

Para realizar este proceso se realizaron las fases de extracción (E), transformación (T) y carga (L), lo cual es fundamental para asegurar la transparencia y trazabilidad del estudio cuantitativo.

En la fase de extracción, en el origen y volumen de datos, se verificó que la extracción inicial arrojó 15,497 registros procedentes del sistema de información estudiantil para estudiantes de primer ingreso comprendidos en los periodos de 2023 a 2025.

El primer paso de esta fase incluyó la anonimización de todos los identificados personales para cumplir con las consideraciones éticas y la ley de protección de datos personales.

Mientras que en la fase de transformación se ejecutó en un flujo de trabajo KNIME, esta plataforma de código abierto que facilita el análisis de los datos. Iniciando con el tratamiento inicial de consistencia y duplicados, para evitar sesgar los resultados, y asegurar que se trabaje con el perfil de primer ingreso.

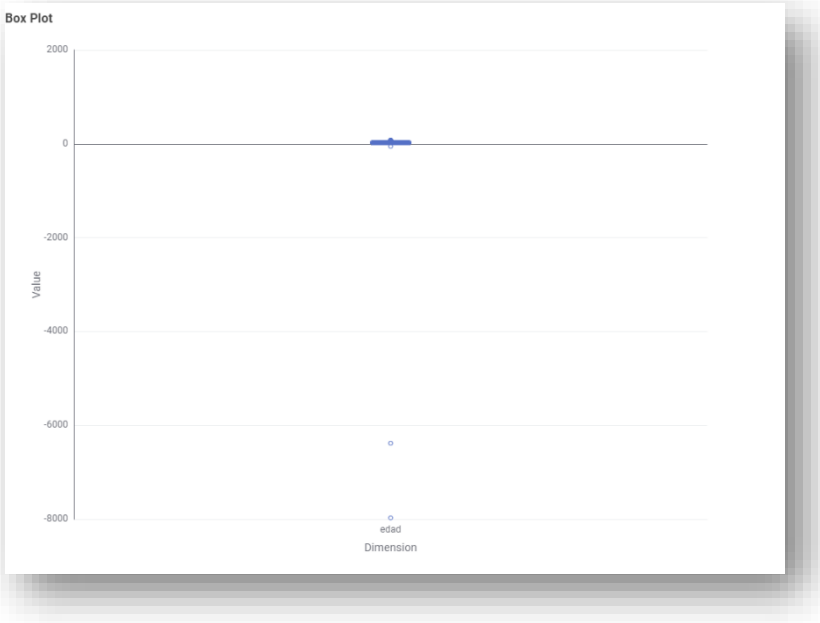


Ilustración 4 - boxplot variable de la edad

El diagrama de caja de la variable edad revela la existencia de valores atípicos extremos, incluyendo edades negativas no plausibles, lo que indica errores en la recopilación o cálculo en los registros originales. Estos valores atípicos distorsionan la escala del gráfico y justificaron la depuración de la variable mediante el recálculo de la edad y la corrección de datos inconsistentes.

Tabla 14 - Resumen de tratamiento de datos faltantes

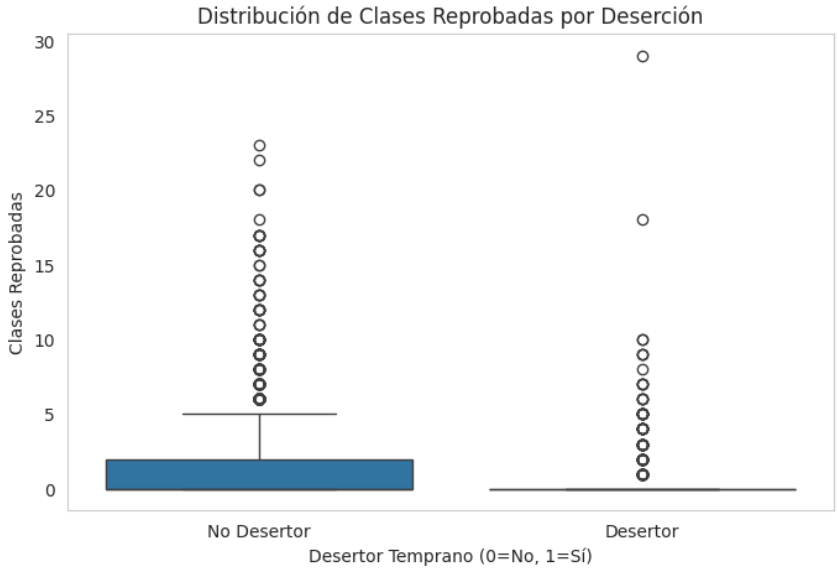
Variable	Porcentaje de nulos	Método de imputación	Justificación
Edad	Aproximadamente 49.0%	Mediana	La mediana es una medida robusta de tendencia central por valores atípicos (como el valor negativo erróneo

			detectado) y es adecuado para la distribución de la edad. Primero se corrigieron los valores negativos a NaN para una imputación más precisa y después se utilizó la mediana.

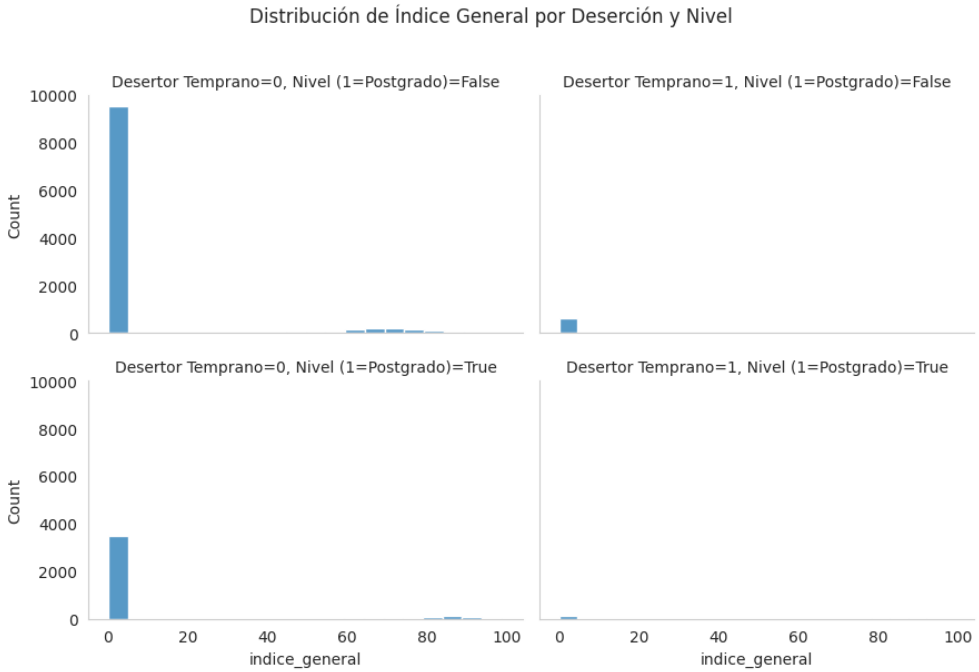
Fuente: Elaboración propia

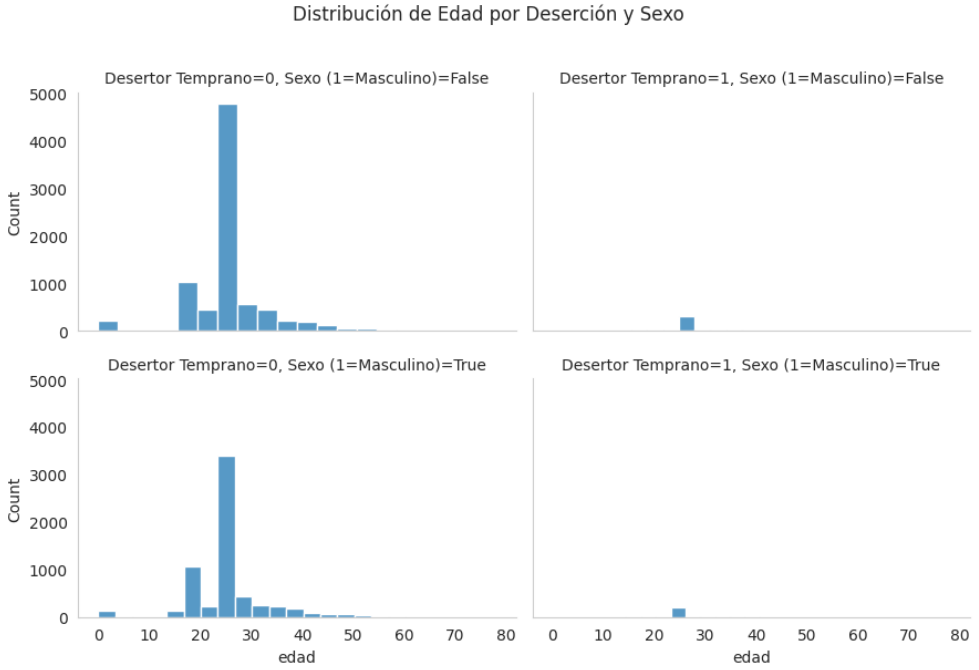
4.1.3 VISUALIZACIÓN DE DATOS

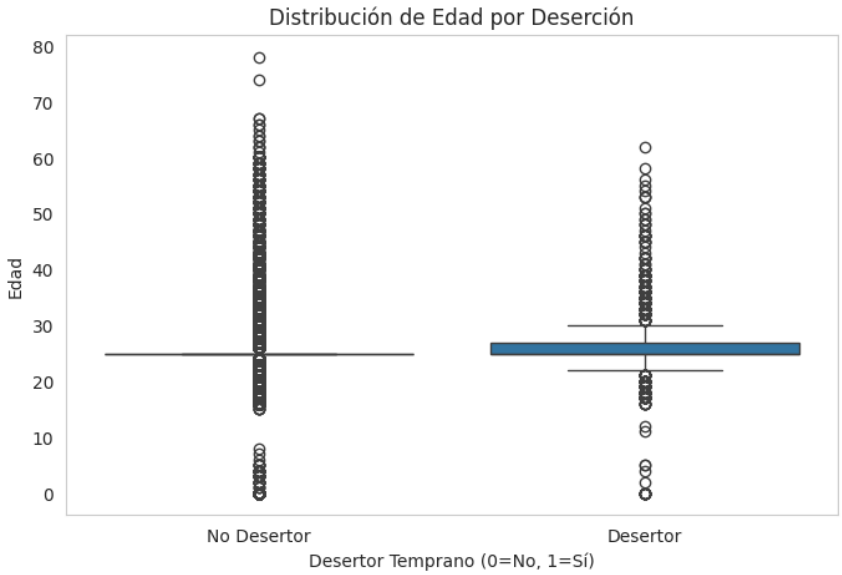
Después de la limpieza, el preprocesamiento y codificación *one-hot* de las variables categóricas, el conjunto total de los datos procesados (*new_train_final*) incluye 350 características (columnas) y 15,497 observaciones (filas). Donde el desertor-temprano es la variable objetivo de la investigación, en la cual examinamos cuántos estudiantes desertaron tempranamente (valor 1) y cuántos no (valor 0). El análisis arrojó que la mayoría de los estudiantes no desertaron tempranamente (alrededor del 94%), lo cual se vuelve fundamental, ya que nos indica que nuestro problema tiene un desbalance de clases, ya que es desproporcional. Esto es clave para cuando construyamos el modelo predictivo.

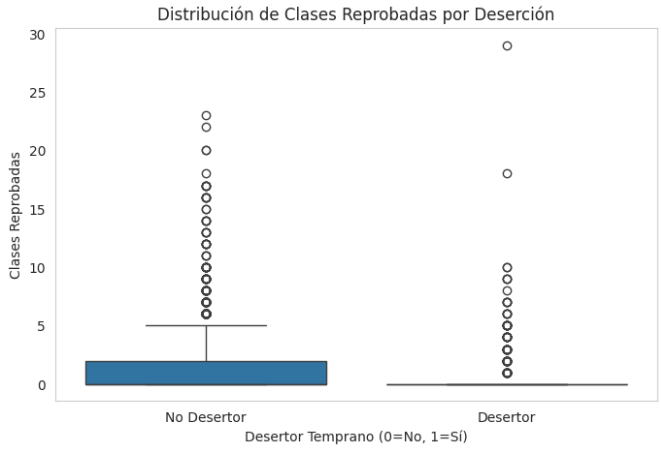


En los gráficos que se exponen a continuación, nos muestran las variables numéricas como; edad, índice general y clases reprobadas; la agrupación de los valores de cada variable y determinar los valores atípicos.

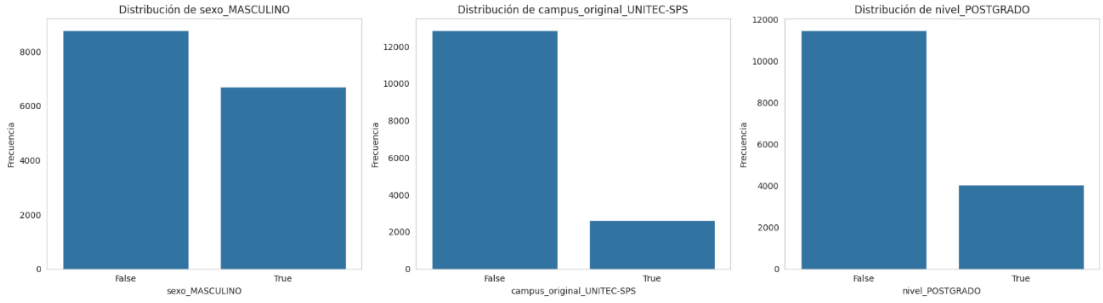

Distribución de Índice General por Deserción y Nivel
Desertor Temprano=0, Nivel (1=Postgrado)=False
Desertor Temprano=1, Nivel (1=Postgrado)=False
Desertor Temprano=0, Nivel (1=Postgrado)=True
Desertor Temprano=1, Nivel (1=Postgrado)=True
Count
indice_general

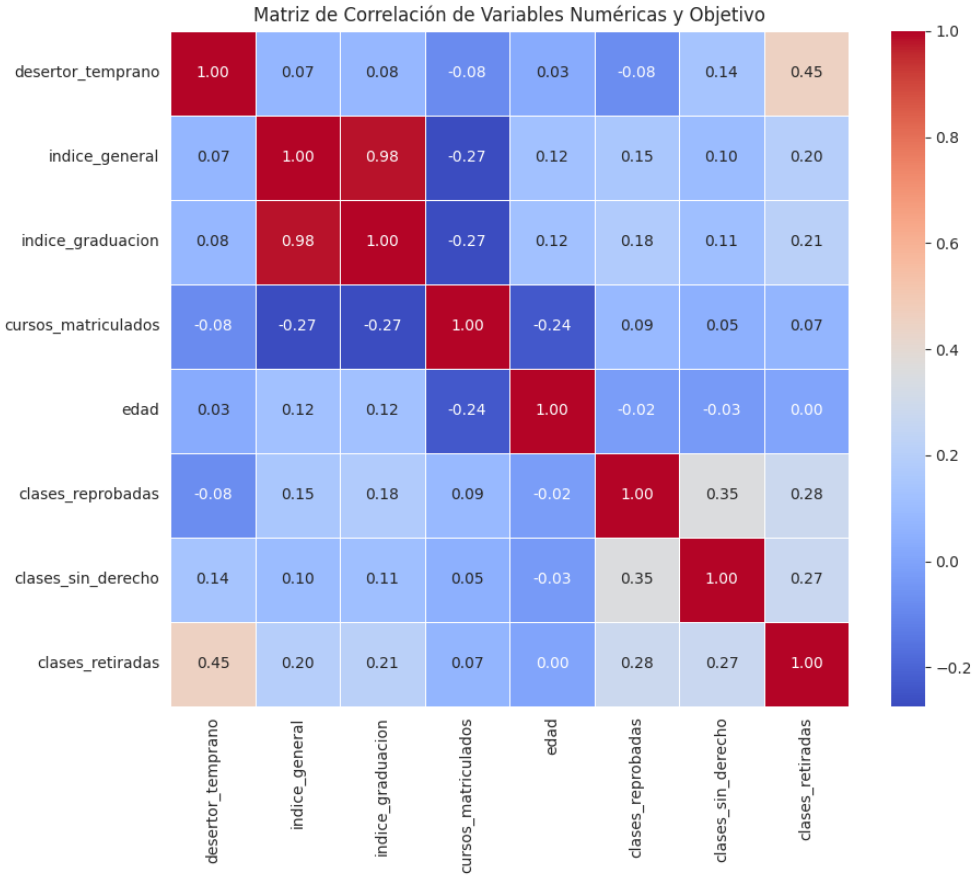

Distribución de Edad por Deserción y Sexo
Desertor Temprano=0, Sexo (1=Masculino)=False
Desertor Temprano=1, Sexo (1=Masculino)=False
Desertor Temprano=0, Sexo (1=Masculino)=True
Desertor Temprano=1, Sexo (1=Masculino)=True
Count
edad





Con relación a las variables categóricas como el género o el nivel, nos indica la cantidad de estudiantes que pertenecen a cada categoría.





La correlación entre las variables numéricas del conjunto de datos y la variable objetivo desertor_temprano se puede observar en la matriz de correlación. En términos generales, las correlaciones son bajas, lo que sugiere que ninguna variable individualmente explica con fuerza la deserción. Algunas relaciones, sin embargo, son destacadas: la deserción tiene una correlación positiva más alta con clases_retiradas (0.45) y el índice general y de graduación tienen correlaciones elevadas entre ellos (0.98), lo que muestra que ambos evalúan aspectos parecidos del desempeño académico. Por otra parte, Las demás variables (edad, cursos_matriculados, clases_reprobadas y clases_sin_derecho) muestran correlaciones débiles con el objetivo, indicando un efecto limitado sobre la deserción en este conjunto de datos.

4.1.4 CONCLUSIONES DEL ANÁLISIS EXPLORATORIO DE DATOS

El análisis exploratorio de datos (EDA) del conjunto de datos de los alumnos ha sido crucial para comprender los elementos vinculados con la deserción escolar temprana. Las conclusiones más importantes son:

Desbalance crítico de la variable objetivo: Se detectó un desequilibrio significativo en la

variable desertor_temprano, con una cifra de alumnos no desertores mucho más alta. Esto pone de relieve la importancia de tener en cuenta métodos específicos de modelado y evaluación para abordar esta desigualdad.

Tratamiento esencial de la edad: La variable edad mostraba una cantidad elevada de valores nulos y valores no normales (negativos). Se llevó a cabo un preprocesamiento exhaustivo, en el que se modificaron los valores negativos a nulos y posteriormente se imputaron estos últimos con la mediana, garantizando así la validez de esta característica.

Factores Clave Asociados a la Deserción:

Género: Se identificó una asociación significativa entre el sexo del estudiante y la deserción temprana, sugiriendo dinámicas de retención diferenciadas por género.

Edad: La edad promedio de los desertores tempranos es significativamente diferente a la de los no desertores, indicando que la etapa vital puede ser un factor.

Rendimiento Académico (índice_general): Existe una diferencia altamente significativa en el índice general entre ambos grupos, confirmando que el bajo rendimiento es un predictor central de la deserción.

Fracaso Académico (clases_reprobadas): El número de clases reprobadas mostró una relación muy fuerte y significativa con la deserción, posicionándose como uno de los predictores más potentes. La acumulación de fracasos académicos es un precursor directo.

Presencia de *Outliers*: Los gráficos de caja (*box plots*) revelaron la existencia de valores atípicos en variables como clases_reprobadas, clases_sin_derecho y clases_retiradas. Estos *outliers*, aunque a veces representan errores, pueden también ser indicadores de casos extremos genuinos con información valiosa.

Implicaciones para el Modelado:

Estos hallazgos sugieren que sexo, edad, índice_general y clases_reprobadas serán variables predictoras fundamentales. Sin embargo, el desbalance de clases requerirá técnicas específicas (como sobremuestreo, submuestreo o el uso de métricas como F1-Score, *Recall* o AUC-ROC) para desarrollar un modelo efectivo y justo. También será importante considerar posibles interacciones no lineales entre estas variables dada la complejidad del fenómeno.

En síntesis, el EDA ha proporcionado una comprensión profunda de los datos y sus relaciones con la deserción temprana, estableciendo una base sólida para la construcción de modelos predictivos robustos y accionables.

4.2 INFORME DEL PROCESO DE RECOLECCIÓN DE DATOS

4.2.1 DESCRIPCIÓN DEL PROCESO

En esta sección se describe el proceso mediante el cual se realizó la recolección de los datos académicos institucionales necesarios para el análisis de la información. Este proceso de ejecutó cumpliendo con los protocolos de acceso y seguridad establecidos por la universidad; ya que ellos fueron los encargados de efectuarlo, a través del departamento de Tecnologías de la información (TI), lo que permitió la obtención de registros fiables. A continuación, se detalla el proceso que llevó a cabo:

Tabla 15 - Fases del proceso de recolección de datos

Etapa	Descripción	Herramientas utilizadas	Duración	Responsables
Recolección de datos	El proceso consistió en extraer y preparar los registros académicos de la universidad, lo que se realizó en un entorno controlado y con el permiso correspondiente de la universidad. Este procedimiento se llevó a cabo utilizando un script *Transact*-SQL, con la base de datos institucional ACAD_DB_PROD. Se aplicó un filtro que solo incluía a los estudiantes de primer ingreso que se matricularon entre 2023 y 2025.	Software de gestión SQL Server Management Studio (SSMS)	10 horas	Departamento de Tecnologías de la Información (TI) de la universidad.
Exportación a formato CVS	Con los datos obtenidos por parte de la institución, se procedió a exportar en formato CSV para su posterior depuración y análisis. Verificando el delimitador la codificación de los caracteres y la integridad del archivo que se generó.	Microsoft Excel		Equipo investigador
Importación de los datos a KNIME	Se cargó el archivo CSV en KNIME para iniciar el procesamiento de los datos y realizar las configuraciones en el nodo de lectura (separador, codificación, tipos de datos	KINME y Python		Equipo investigador

	iniciales) y verificar que el total de los registros coincida con el total de los estudiantes que asciende a 15,497 registros válidos.			
Limpieza y preparación de los datos	Se realizó el respectivo tratamiento de los valores nulos, a través del mapa de calor y una vez detectadas las inconsistencias se procedió con la debida corrección.	KINME y Python		Equipo investigador
Transformación y codificación de las variables	Se codificaron los datos de acuerdo con los requerimientos necesarios para su análisis.	KINME y Python		Equipo investigador
Anonimización de los datos	Se eliminaron los identificadores personales de los estudiantes, garantizando la confidencialidad de los estudiantes.	KINME y Python		Equipo investigador
Generación de la base empírica	Se consolidaron los datos limpios depurado y transformado en un archivo final para su respectivo análisis estadístico y modelado.	KINME y Python		Equipo investigador

Fuente: Elaboración propia

4.2.2 PARTICIPANTES O FUENTES DE INFORMACIÓN

Los datos proceden de los registros académicos de 15,497 estudiantes de primer ingreso entre los años 2023 y 2025.

Estos registros comprenden información académica, demográfica y administrativa de los estudiantes.

Tabla 16 Perfil descriptivo de la población estudiantil

Categoría	Descripción
Distribución de género (%)	53% Femenino y 47% masculino
Edad promedio	La edad promedio oscila entre 18 a 22 años
Modalidad	Presencial y semipresencial
Carreras predominantes	Ingenierías, Administración, Contaduría y Psicología

Fuente: Elaboración propia

El uso de datos secundarios institucionales garantiza la validez, representatividad y confiabilidad de la información (Hernández Sampieri & Mendoza Torres, 2018)

4.2.3 INSTRUMENTOS UTILIZADOS

Los instrumentos utilizados para esta investigación se utilizaron de carácter documental, tecnológico y analítico, en una primera instancia estos se emplearon bases de datos académicas

institucionales, las cuales proporcionan información estructurada sobre el rendimiento académico, historial de asignaturas, periodos cursados y características de los estudiantes de primer ingreso.

Los instrumentos tecnológicos utilizados son los siguientes:

1. **Python (Pandas, *Numpy*, *scikit-learn*, *matplotlib*)**

Se utilizo para la limpieza, procesamiento y depuración de los datos, así como para la construcción de modelos predictivos, permitió manejar grandes volúmenes de información y aplicar técnicas de minería de datos de forma reproducible.

2. ***Knime* plataforma de analítica de datos**

Este software nos permitió realizar los flujos de trabajo visuales para la transformación, imputación, normalización y selección de variables, así como realizar la ejecución y evaluación de algoritmos de clasificación, su uso está ampliamente validado en investigaciones de *Educational Data Mining*.

3. **Documentación institucional**

Se emplearon manuales académicos, lineamientos internos, reglamentos y reportes de retención estudiantil para comprender el contexto y validar las definiciones operativas

Los instrumentos utilizados fueron de carácter documental y tecnológico (softwares).

Se seleccionaron estos instrumentos debido a que se basan en investigaciones previas sobre análisis educativo y aprendizaje automático, como señalan (Cristobal Romero & Ventura, 2020) quien ha documentado ampliamente su validez en estudios de predicción de abandono académico.

Ya que se empleó la base de datos académicas institucional, las cual proporcionó la información sobre el desempeño académico y características de los estudiantes. En cuanto a los softwares, se utilizó Python y KNIME, para la limpieza, procesamiento, depuración y transformación de los datos; así como para la construcción y ejecución de modelos predictivos orientados a identificar patrones asociados a la deserción estudiantil, permitiendo de esta manera, garantizar la calidad y solidez del análisis realizado.

Estos instrumentos fueron seleccionados por su uso validado en la minería de datos educativa y su respaldo en investigaciones previas (Cristobal Romero & Ventura, 2020)

4.2.4 DIFICULTADES ENCONTRADAS

Dificultad	Descripción	Estrategia de resolución
Inconsistencias en los datos institucionales	Se identificaron errores tipográficos en nombres de carreras, códigos duplicados, asignaturas mal registradas y valores faltantes en variables clave como la edad. Estas inconsistencias dificultan la integración y análisis de la información.	Se desarrollo un script de normalización de datos en Python y se aplicaron varios nodos de limpieza en *Knime*, los valores faltantes fueron tratados mediante imputación por mediana, garantizando menor distorsión en la distribución.
Retraso en la autorización de acceso a las bases de datos.	La autorización por parte de institución se retrasó una semana, ya que debía escalar diferentes permisos.	Se tuvo que ajustar el cronograma, evitando afectar la modelización posterior de los datos.
Limitación de variables no académicas	Se encontró la total ausencia de factores socioeconómicos y psicológicos de los estudiantes asociados al desempeño y deserción académica.	Se documentó como una limitación la ausencia de estos factores, además, se recomendó ampliar las fuentes futuras para un análisis más integral de las causas que inciden en la deserción estudiantil.

Fuente: Elaboración propia

4.2.5 CONSIDERACIONES ÉTICAS

Se usaron los principios éticos de la investigación científica y la Ley de Protección de Datos Personales (Decreto No. 25-2017).

Medidas aplicadas:

- Confidencialidad: se determinó ocultar todos los identificadores personales de los estudiantes con el propósito de proteger la privacidad de cada uno de ellos y garantizar la confidencialidad de la información; en cumplimiento de los principios éticos de la investigación.
- Consentimiento institucional: se obtuvo autorización escrita de la universidad para el acceso y manejo de la información académica, denotando un proceso dentro del marco legal y ético.
- Uso exclusivo: los datos se emplearon únicamente para fines académicos, asegurando que los mismos no serán empleados para ningún otro propósito que no sea el investigativo.
- Transparencia y trazabilidad: se documentó cada fase del proceso técnico, así como las condiciones, límites y responsabilidades asociadas al uso de los datos. De igual

manera, los procedimientos de trazabilidad a través del registro de cada etapa garantizan que todas las acciones se realizaron con base en lo establecido por la institución.

- Cumplimiento ético: se aplicaron las recomendaciones de (AERA, 2011) sobre la integridad y privacidad en el manejo de datos educativos, con el fin de preservar completamente los derechos y la dignidad de los estudiantes.

Con el cumplimiento de estos principios se asegura la validez científica y la responsabilidad social del estudio (Resnik, 2024).

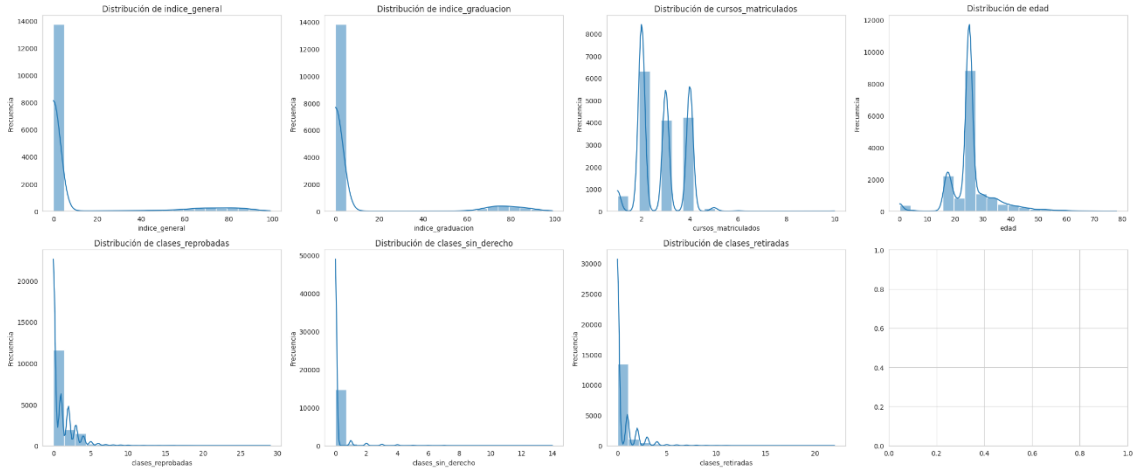
4.3 RESULTADOS Y ANÁLISIS DE LAS TÉCNICAS APLICADAS

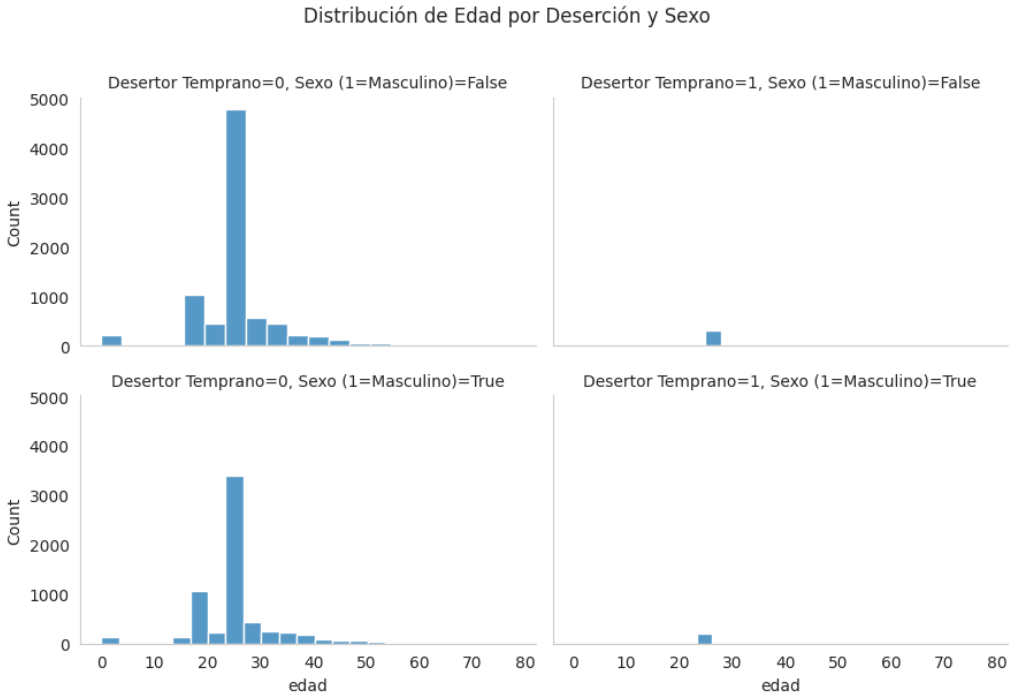
4.3.1 RESULTADOS CUANTITATIVOS

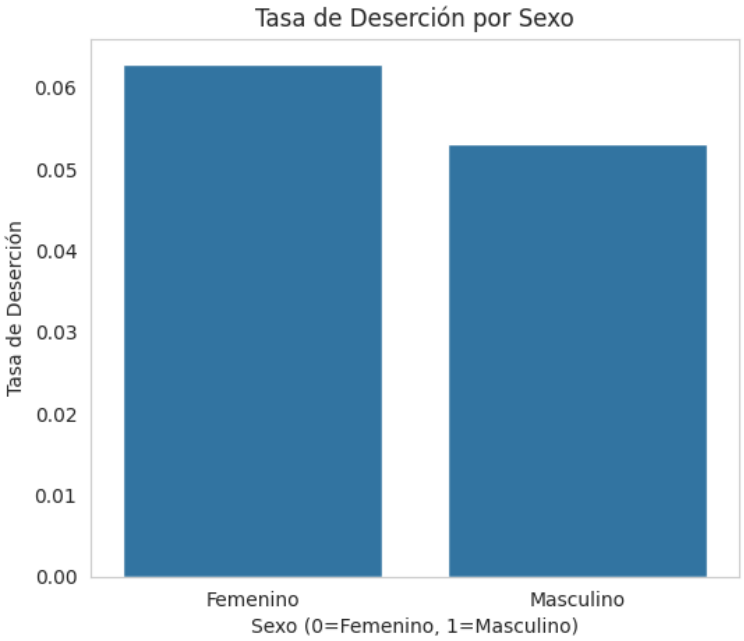
4.3.1.1 PRESENTACIÓN DE DATOS

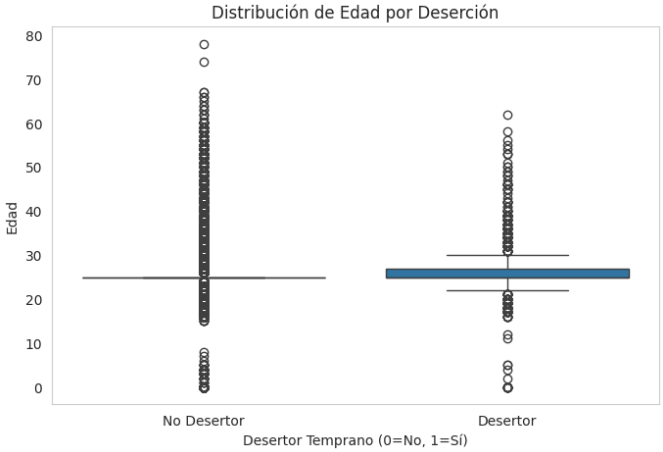
Para el análisis cuantitativo se procesaron varios registros académicos de los estudiantes de primer ingreso que tiene al menos una clase reprobada, los resultados han sido organizados a través de tablas de distribución, gráficos de tendencia y diagramas comparativos, lo que ha permitido identificar comportamientos recurrentes en la población estudiada.

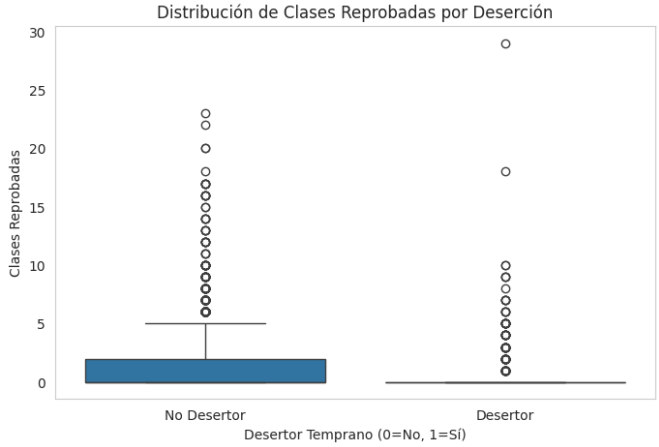
La presentación de los datos se estructuró de forma descriptiva, permitiendo observar variables como:




Distribución de Edad por Deserción y Sexo
Desertor Temprano=0, Sexo (1=Masculino)=False
Desertor Temprano=1, Sexo (1=Masculino)=False
Desertor Temprano=0, Sexo (1=Masculino)=True
Desertor Temprano=1, Sexo (1=Masculino)=True
Count
edad
0
1000
2000
3000
4000
5000
0
10
20
30
40
50
60
70
80


Tasa de Deserción por Sexo
Tasa de Deserción
0.00
0.01
0.02
0.03
0.04
0.05
0.06
Femenino
Masculino
Sexo (0=Femenino, 1=Masculino)





4.3.1.2 DESCRIPCIÓN DE HALLAZGOS

Los resultados preliminares sugieren:

- Que hay un incremento en la cantidad de estudiantes con materias reprobadas en distintos periodos académicos.
- La concentración de reprobaciones es en materias básicas o introductorias.
- Existen diversas variaciones notables entre cohortes y modalidades de estudio.
- La relación entre carga académica alta y mayores tasas de riesgo académico.

4.3.1.3 RELACIÓN CON LOS OBJETIVOS

Varios resultados se relacionan directamente con el objetivo principal de este estudio, el cual es analizar el riesgo de deserción en estudiantes de primer ingreso con materias reprobadas.

El análisis descriptivo nos permitió determinar ciertas tendencias relevantes que ayudan a alimentar la etapa posterior de modelado predictivo, ayudando a responder a preguntas clave como:

- ¿Cuáles patrones de reprobación se asocian con mayor riesgo de abandono?
- ¿Existen cohortes o asignaturas que ayudan a concentrar la vulnerabilidad académica? ¿Qué comportamientos son comunes en estudiantes que se encuentran en riesgo académico?

4.3.1.4 ANÁLISIS ESTADÍSTICO

El análisis estadístico se realizó con el propósito de describir el desempeño académico de los estudiantes y establecer si existen diferencias significativas entre los estudiantes que han desertado y los que se encuentran activos. Para lo cual, se hizo uso de estadísticas descriptivas, prueba t para grupos independientes y prueba de asociación chi cuadrado. Los resultados fueron elaborados en KNIME con los nodos, Independent Samples t-test y Chi-Square Test.

Se usa la prueba t-test para grupos independientes (Independent *groups* t-test), ya que permite hacer una comparación estadística entre dos grupos independientes (desertores y no desertores) en variables académicas continuas. Esta prueba es la idónea para determinar si el rendimiento es diferente entre los dos grupos y proporciona información para analizar la deserción.

Independent Groups Statistics

Confidence Interval (CI) Probability: 95.0%

Differences are reported of the groups: 1 - 0

	Variance Assumption	t	df	p-value (2-tailed)	Mean Difference	Standard Error Difference	CI (Lower Bound)	CI (Upper Bound)
indiceGeneral	Equal variances assumed	18.1227	15,495	1.19E-72	7.4176	0.4093	6.6153	8.2199
indiceGeneral	Equal variances not assumed	15.877	6,103.5549	1.17E-55	7.4176	0.4672	6.5017	8.3334
indiceGraduacion	Equal variances assumed	18.6766	15,495	5.37E-77	8.1374	0.4357	7.2834	8.9914
indiceGraduacion	Equal variances not assumed	16.1526	5,989.4316	1.76E-57	8.1374	0.5038	7.1498	9.125
cursosMatriculados	Equal variances assumed	-28.1786	15,495	2.05E-170	-0.4595	0.0163	-0.4915	-0.4276
cursosMatriculados	Equal variances not assumed	-29.8295	8,638.1131	3.53E-186	-0.4595	0.0154	-0.4897	-0.4293
clases_reprobadas	Equal variances assumed	35.4407	15,495	1.30E-264	1.1263	0.0318	1.064	1.1886
clases_reprobadas	Equal variances not assumed	26.9391	5,097.1509	1.48E-149	1.1263	0.0418	1.0443	1.2083
clases_sin_derecho	Equal variances assumed	12.0686	15,495	2.18E-33	0.13	0.0108	0.1089	0.1511
clases_sin_derecho	Equal variances not assumed	9.0813	5,042.7032	1.51E-19	0.13	0.0143	0.1019	0.158
clases_retiradas	Equal variances assumed	31.3835	15,495	1.17E-209	0.6533	0.0208	0.6125	0.6941
clases_retiradas	Equal variances not assumed	22.2604	4,758.1603	1.64E-104	0.6533	0.0293	0.5958	0.7109

Ilustración 5 - Prueba t de grupos independientes

Se utilizó la t-test de grupos independientes para comparar el rendimiento académico entre desertores y no desertores. Estos resultados demostraron diferencias estadísticamente significativas ($p < 0.001$) para todas las variables que se han estudiado, tales como; índice académico, clases matriculadas, clases reprobadas y clases retiradas.

Estas diferencias en los promedios demuestran que los dos grupos tienen perfiles académicos diferentes y, por lo tanto, estas variables son unos buenos indicadores de riesgo de deserción.

Para determinar la asociación entre variables categóricas y la deserción escolar se utilizaron tablas de contingencia (*Crosstab*) y la prueba de independencia Chi-cuadrado. Este proceso permitió establecer si la distribución de desertores y no desertores difería en términos de características académicas o demográficas.

Tabla 16 – Resumen de la prueba Chi-cuadrado

Variable	DF	X^2	p-*value*
campus_original	2	1,398.67	1.92E-304
periodo	7	195.3149	1.13E-38
desertor_temprano	1	1,889.62	0
desertor_total	1	12,027.32	0
bajo_índice	1	274.8554	9.93E-62
Sexo	2	23.3921	8.33E-06
permiso_matrícula_por_periodo	1	103.6538	2.41E-24
nivel	2	242.4912	2.21E-53
tipo_ingreso_1	0	0	NaN
tipo_ingreso_2	4	655.7033	1.36E-140
códigoCarrera	172	2,650.79	0
nombreCarrera	147	2,604.27	0
facultad	5	1,159.96	1.39E-248

Fuente: Elaboración propia

Se efectuó la prueba de Chi-cuadrado a las variables categóricas con el fin de determinar su asociación con la deserción estudiantil. La mayoría de las variables arrojaron valores de $p < 0.05$, lo que indica asociaciones estadísticamente significativas. Esto quiere decir que variables como campus, nivel, facultad, forma de ingreso y permisos de matrícula se asocian a mayor o menor deserción. Las variables no significativas estadísticamente no se asociaron a la deserción. Estos resultados guían la elección de predictores para los modelos posteriores.

Según Hernández Sampieri & Mendoza Torres (2018), la aplicación de las pruebas estadísticas en las investigaciones mixtas fortalece la validez de los hallazgos, lo cual permite un mayor rigidez en la interpretación de los datos.

4.3.2 ANÁLISIS CUALITATIVO

4.3.2.1 CATEGORÍAS O TEMAS EMERGENTES

El análisis cualitativo se realizó mediante la codificación temática, con coherencia usando el enfoque propuesto, las categorías emergentes fueron:

- Factores personales: son relevantes ya que permiten identificar patrones que van más allá del análisis del desempeño académico medido a través de las calificaciones. Entre ellos tenemos las características individuales de los estudiantes, como; la motivación, la gestión de tiempo, la adaptación al entorno universitario; que también pueden influir en el riesgo de deserción académica.
- Factores académicos: estos factores son fundamentales para la predicción, ya que permiten identificar patrones de desempeño que mejoran la precisión de las estimaciones y brindan un aporte valioso para la implementación de intervenciones institucionales desde el acompañamiento académico. Entre ellos están, la carga de trabajo, dificultad de contenido, interacción entre docente y estudiante; las cuales juegan un papel primordial en la permanencia de los estudiantes.
- Factores institucionales: los cuales están asociados con la calidad de los servicios académicos y administrativos, entre los más relevantes están; el acompañamiento y orientación académica, la accesibilidad a tutorías, la modalidad de clase. Estos son esenciales, ya que permiten comprender cómo el entorno académico impacta en el riesgo de deserción académica. Estas categorías nos ayudan a comprender elementos no cuantificables que inciden en el rendimiento y permanencia académica.

Estas categorías nos ayudan a comprender elementos no cuantificables que inciden en el rendimiento y permanencia académica.

4.3.2.2 CITAS O EJEMPLOS

Se utilizó extractos representativos de respuestas abiertas los cuales reflejan percepciones comunes, tales como:

- “Los cursos de matemática son muy difíciles y no hay suficiente apoyo”
- “Me costó adaptarme porque yo trabajo y estudio al mismo tiempo”.

4.3.2.3 INTERPRETACIÓN Y RELACIÓN CON EL MARCO TEÓRICO

Este análisis reveló que la deserción no depende únicamente del rendimiento académico, sino que también de los factores personales, académicos e institucionales; los cuales inciden en el desempeño académico; ya que la parte motivacional, emocional y contextual, impactan significativamente.

Tal cómo se refleja en el análisis PESTEL los aspectos económicos tienen una influencia en el porcentaje de deserción en universidades privadas, por lo cual son variables importantes de analizar con el fin de obtener una información más fiable sobre las causas de la deserción.

La aplicación de herramientas de análisis predictivo permite tomar decisiones más acertadas y oportunas para brindar un eficiente acompañamiento a los estudiantes, ya que este fue identificado como un factor crítico, reforzando la importancia de estrategias institucionales proactivas. Este es elemento clave y que genera una ventaja competitiva al contar con una gestión más eficiente de las bases de datos y registros de los estudiantes.

4.3.2.4 TRIANGULACIÓN DE DATOS

La triangulación de datos cuantitativos, cualitativos permitió:

- Validar patrones numéricos mediante testimonios y percepciones estudiantiles.
- Contrastar la realidad institucional con hallazgos estadísticos.
- Aumentar la robustez interpretativa de la investigación.

4.4 ANÁLISIS INFERENCIAL Y MODELOS APLICADOS

La siguiente sección, trata del objetivo principal de esta investigación: determinar las asociaciones estadísticas entre las principales variables predictoras y verificar los modelos predictivos para calcular el riesgo de deserción estudiantil en estudiantes. Esta parte representa el

diseño no experimental de tipo predictivo y el enfoque cuantitativo de la investigación, usando técnicas de minería de datos y aprendizaje automático.

Para lograr esto, la sección alcanzará el entrenamiento y la evaluación de los principales algoritmos de clasificación: *Random Forest, K Nearest Neighbor, Decision Tree y Regression Predictor.*

El análisis inferencial intentará medir la magnitud de la relación entre la variable objetivo ('desertor') y las principales variables predictoras encontradas en el AED, tales como: índice general, cantidad de clases reprobadas, edad y sexo. etc.

La validación de los modelos es fundamental, por la naturaleza del problema y por el desbalance extremo que se observa en la variable objetivo ('desertor'). Para ello, se utilizarán métricas estrictas como exactitud, precisión, exhaustividad, F1 y AUC-ROC, y se aplicará la validación cruzada para garantizar la robustez y prevenir el sobreajuste. Estos resultados podrán establecer la exactitud predictiva y anticipatoria de los modelos planteados, comparándolos con el seguimiento académico habitual

4.4.1 ANÁLISIS INFERENCIAL

Para confirmar estadísticamente las relaciones encontradas en la etapa exploratoria y establecer si las diferencias encontradas entre los grupos (desertores y no desertores) se pueden generalizar a la población y no son producto del azar, se utilizaron pruebas de hipótesis paramétricas y no paramétricas. la estadística inferencial sirve para estimar parámetros y probar hipótesis, basándose en la distribución muestral.

A continuación, se muestran los resultados de las pruebas de significancia estadística con un nivel de confianza del 95% (α=0.05):

Diferencia de medias (t de Student)

Para determinar si existen diferencias significativas en el rendimiento académico entre desertores y desertores se utilizó la prueba t para muestras independientes. Esta prueba es adecuada para comparar si dos grupos son diferentes en sus medias en una variable cuantitativa.

Variable Crítica "Clases Reprobadas": El análisis reveló una diferencia altamente significativa ($p < 0.001$, valor exacto: 1.30E-264). Los desertores tienen un promedio de 2.55 cursos reprobados, en comparación con 0.94 de los retenidos. Como el valor p es menor que el

nivel de significancia (0.05), se rechaza la hipótesis nula (H0). lo que verifica que la acumulación de reprobaciones sí es un elemento diferenciador real y no aleatorio entre los grupos.

Asociación de variables categóricas (prueba chi-cuadrada)

Para verificar si la deserción es independiente de variables institucionales como "Campus" o "Facultad", se aplicó la prueba de Chi-cuadrada de Pearson ($\chi 2$). esta prueba es una medida de diferencia entre las frecuencias observadas y las frecuencias esperadas, que determina si existe asociación estocástica entre dos variables nominales.

Dependencia institucional: Los resultados revelaron una asociación estadísticamente significativa ($p<0.05$) para las variables "Campus de Origen" y "Facultad". Esto quiere decir que la deserción no se distribuye de manera homogénea, hay campus y facultades en los que se da estadísticamente más deserción de lo esperado, justificando así la inclusión de estas variables categóricas como predictoras en los modelos de aprendizaje automático.

En cambio, otras variables como "Tipo de Ingreso 1" no alcanzaron significancia estadística ($p > 0.05$) o mostraron valores NaN (Not a Number) en algunas categorías, indicando independencia o falta de variabilidad suficiente para ser considerados predictores confiables en este modelo.

4.4.2 MODELOS APLICADOS

En esta parte se explican los modelos predictivos desarrollados para identificar los estudiantes en riesgo de deserción. Después del análisis inferencial, se eligieron algoritmos comunes en estudios de educación y minería de datos por su capacidad para trabajar con algoritmos complejos y variables heterogéneas.

Se entrenaron los modelos *Gradient Boosted Trees Random Forest, Decision Tree, K-Nearest Neighbors* y Regresión Logística, usando *Stratified K-Fold Cross Validation* para garantizar una evaluación justa ante el desbalance de clases. Cada modelo se midió con métricas comunes (exactitud, *recall,* precisión, F1-score y *Cohen's kappa*) y matrices de confusión creadas en *KNIME*. Estos resultados se pueden utilizar para comparar algoritmos y ver cuáles son mejores para predecir la deserción estudiantil.

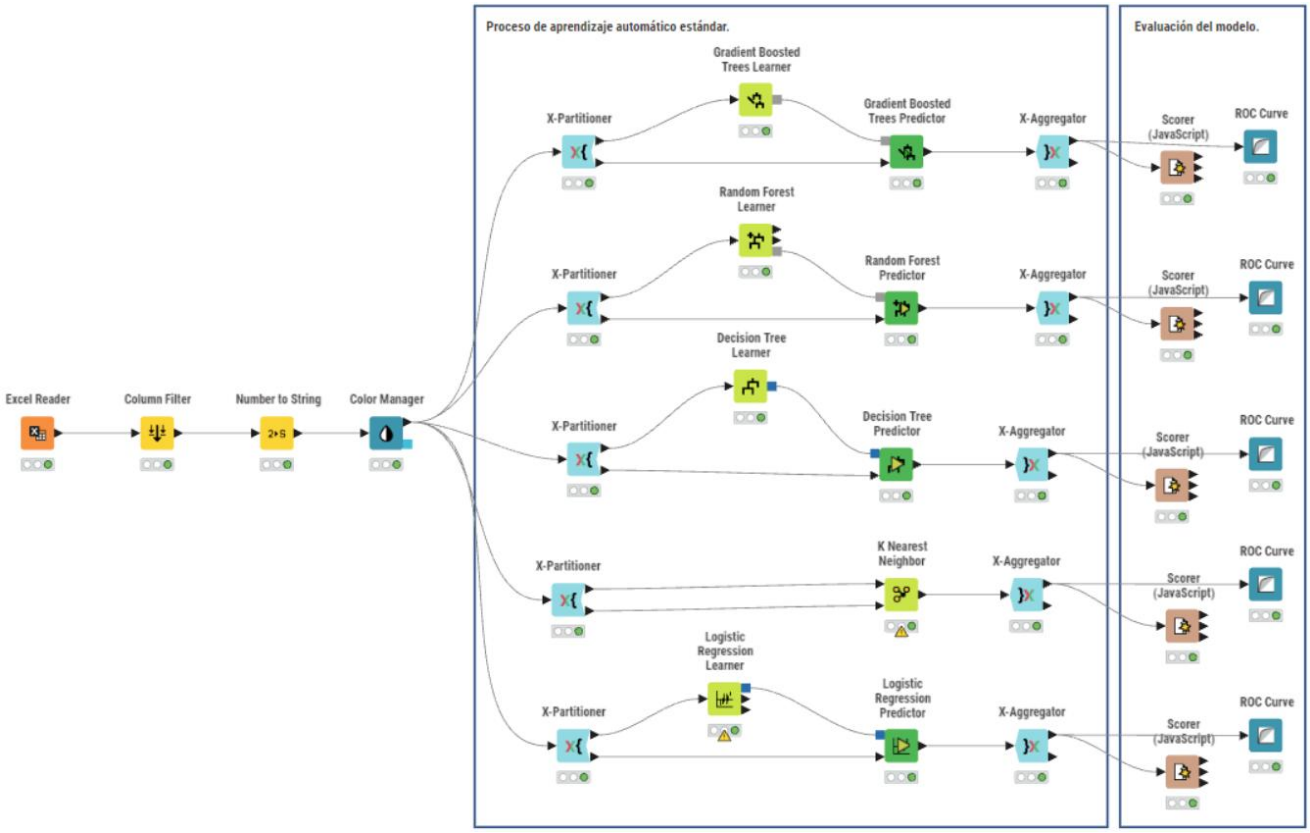


Ilustración 6 - Modelos predictivos para la deserción estudiantil.

La ilustración 7 representa la arquitectura completa del proceso de aprendizaje automático en KNIME para predecir el riesgo de deserción estudiantil. El proceso se inicia con la lectura del archivo en formato Excel con Excel Reader, luego se prepara y limpia la información, filtrando variables (Column Filter), convirtiendo (*Number to String*) y estandarizando visualmente (*Color Manager*). Luego, los datos procesados se dirigen a varias ramas paralelas del proceso, cada una para un modelo predictivo diferente: *Gradient Boosted Trees, Random Forest, Decision Tree, K-Nearest Neighbors,* Regresión Logística.

Cada modelo hace uso de un nodo *X-Partitioner*, el cual crea automáticamente las particiones para validación cruzada, asegurando la reproducción del mismo proceso de entrenamiento en un esquema estandarizado. Se utiliza mejor la validación cruzada como una herramienta metodológica robusta para evaluar la estabilidad y la capacidad de generalización de los modelos predictivos, superando las limitaciones de una simple partición de datos en entrenamiento y prueba. Según (Acito, 2023), el uso de métodos de partición *k-fold* es fundamental para asegurar que los modelos, tales como los árboles de decisión, tengan la capacidad de predecir correctamente sobre datos no vistos (*unseen data*), evitando así que el algoritmo "descubra"

asociaciones espurias que son meros artefactos de un conjunto de datos específico

A continuación, cada algoritmo entrena su modelo a través de los nodos Learner y realiza predicciones con los nodos Predictor. Las salidas se combinan en un X-*Aggregator*, que reúne los resultados de todos los pliegues y proporciona métricas robustas y generalizables.

Finalmente, cada modelo se somete a dos métricas de evaluación: (1) *Scorer* (JavaScript), que genera la matriz de confusión, métricas de precisión, sensibilidad, especificidad y F1; y (2) Curva ROC, para evaluar la capacidad discriminativa del modelo a través del AUC.

Este flujo unificado somete a todos los modelos bajo las mismas condiciones de entrenamiento, validación y prueba, permitiendo una comparación metodológicamente consistente y técnicamente sólida entre algoritmos.

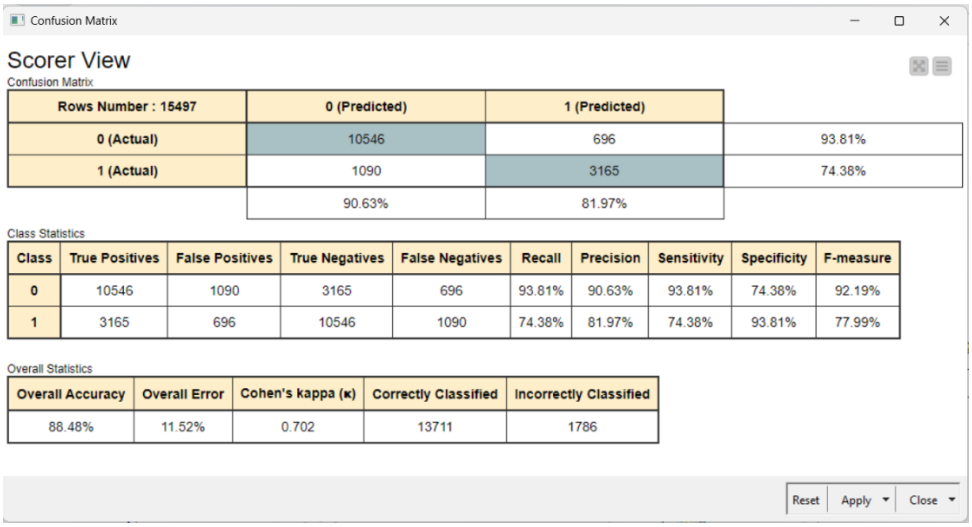
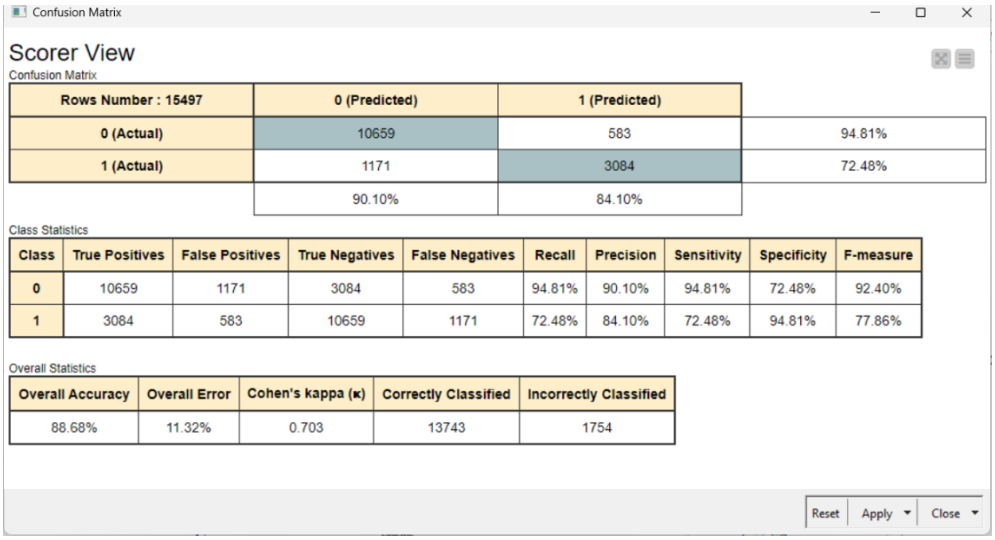


Ilustración 7 - Matriz de confusión del modelo Gradient Boosted Trees

El modelo *Gradient Boosted Trees* mostró un desempeño robusto en la predicción del riesgo de deserción estudiantil. La matriz de confusión evidenció una alta capacidad discriminativa, clasificando correctamente a 13,711 estudiantes (88.48% del total). El modelo identificó adecuadamente a los estudiantes que no desertaron (Clase 0) con una sensibilidad del 93.81%, mientras que para la clase de deserción (Clase 1) alcanzó una sensibilidad del 74.38%, reflejando una mejora sustancial frente a métodos tradicionales. La precisión de la clase positiva

(81.97%) y el F-*measure* de 77.99% indican un equilibrio adecuado entre falsos positivos y falsos negativos. El coeficiente *Cohen's Kappa* = 0.702 confirma un nivel de acuerdo "substancial" entre predicción y realidad, demostrando que el modelo supera ampliamente el azar y es estadísticamente confiable para apoyar decisiones institucionales de intervención temprana.


Confusion Matrix

Scorer View

Confusion Matrix

Rows Number : 15497	0 (Predicted)	1 (Predicted)	
0 (Actual)	10659	583	94.81%
1 (Actual)	1171	3084	72.48%
	90.10%	84.10%	

Class Statistics

Class	True Positives	False Positives	True Negatives	False Negatives	Recall	Precision	Sensitivity	Specificity	F-measure
0	10659	1171	3084	583	94.81%	90.10%	94.81%	72.48%	92.40%
1	3084	583	10659	1171	72.48%	84.10%	72.48%	94.81%	77.86%

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
88.68%	11.32%	0.703	13743	1754

Reset Apply Close

Ilustración 8 - Matriz de confusión del modelo Random Forest

El modelo Random Forest rindió de manera satisfactoria y constante en la predicción del riesgo de deserción estudiantil. La matriz de confusión indica que el algoritmo clasificó correctamente a 13,743 estudiantes, para una precisión global del 88.68%, lo que demuestra su estabilidad ante datos altamente heterogéneos. Sensibilidad de la clase no deserta (0) fue alta (94.81%), demostrando una buena identificación de estudiantes desertores. Para la deserción de clase (1), el modelo alcanzó una sensibilidad del 72.48%, superando el rendimiento de las formas tradicionales de seguimiento manual. Con una exactitud del 84.10% y un F-*measure* de 77.86%, se logra un buen balance entre falsos positivos y falsos negativos. Finalmente, el coeficiente de *Cohen's* Kappa = 0.703 muestra una concordancia "sustancial", lo que verifica la fiabilidad del modelo y su potencial uso institucional para alertas tempranas.

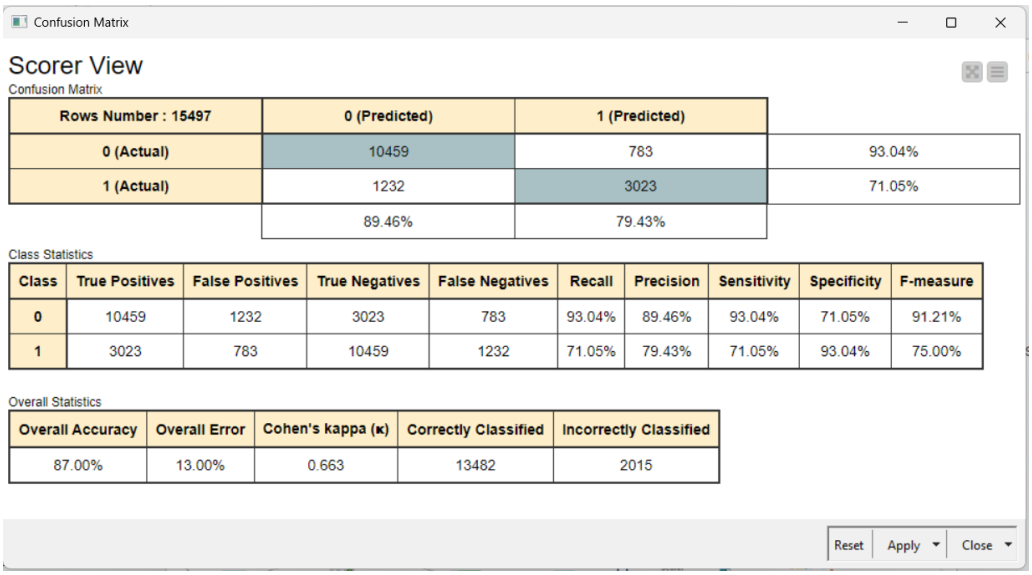


Ilustración 9 - Matriz de confusión del modelo Decision Tree

El modelo *Decision Tree* logró una exactitud global de 87.00%, clasificando correctamente 13,482 de los 15,497 casos probados. La matriz de confusión muestra un rendimiento desbalanceado entre clases: la clase No desertó (0) tiene un *recall* de 93.04%, lo que demuestra la capacidad del modelo para reconocer estudiantes activos. Pero para la clase Desertó (1) el *recall* se reduce a 71.05%, mostrando menor sensibilidad para identificar casos de deserción. La exactitud para las clases 0 y 1 fue de 89.46% y 79.43%, respectivamente, y el índice *kappa* (0.663) muestra una concordancia sustancial entre lo predicho y lo real. Si bien el árbol de decisión funciona decentemente, es mucho menos efectivo identificando estudiantes en riesgo que modelos más poderosos como *Gradient Boosted Trees o Random Forest*.

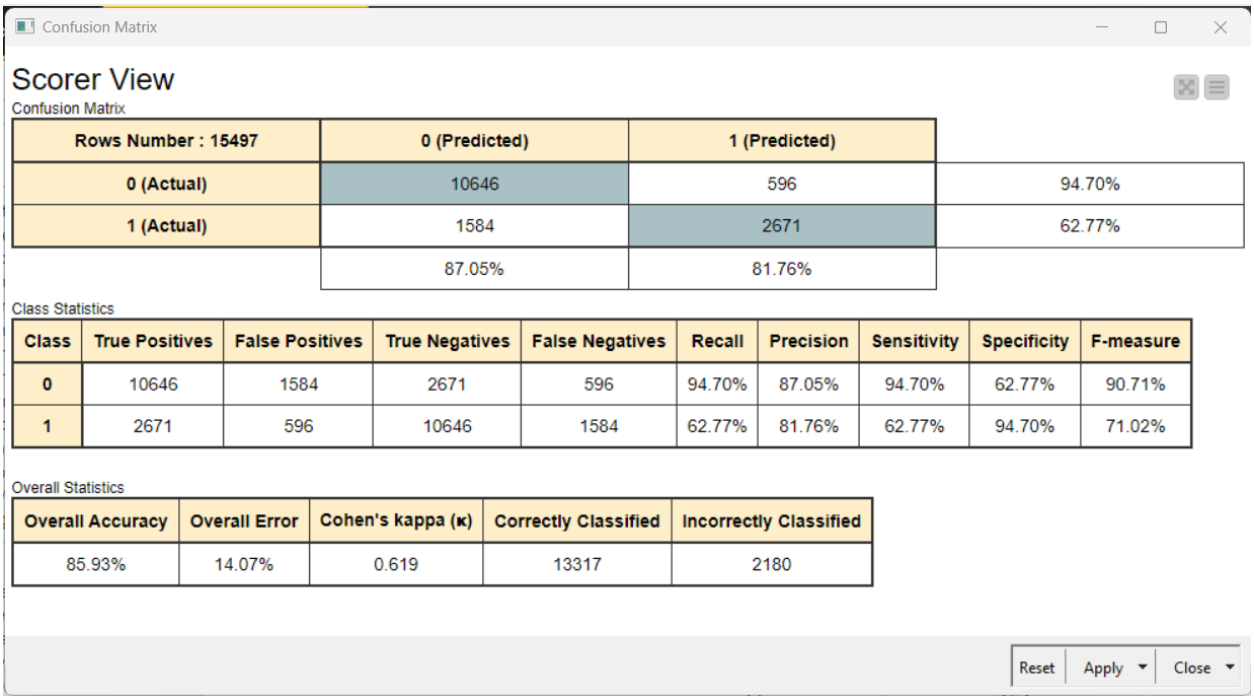


Ilustración 10 - Matriz de confusión del modelo K-Nearest Neighbor

El modelo KNN arrojó una precisión global de 85.93%, con 13,317 instancias clasificadas correctamente. La matriz de confusión muestra un desempeño desbalanceado entre las clases: para la clase No desertó (0), el modelo logró un *recall* de 94.70%, mostrando una buena capacidad para reconocer estudiantes activos. Por el contrario, para la clase Desertó (1) el rendimiento fue mucho menor, con un *recall* de 62.77%, lo que refleja problemas para identificar casos de deserción. La exactitud fue de 87.05% para la clase 0 y 81.76% para la clase 1; el índice *kappa* (0.619) indica una concordancia moderada entre lo predicho y lo real. Estos resultados indican que, si bien KNN se desempeña de manera aceptable en general, su capacidad discriminativa para el grupo en riesgo es menor a la de modelos basados en árboles o ensambles.

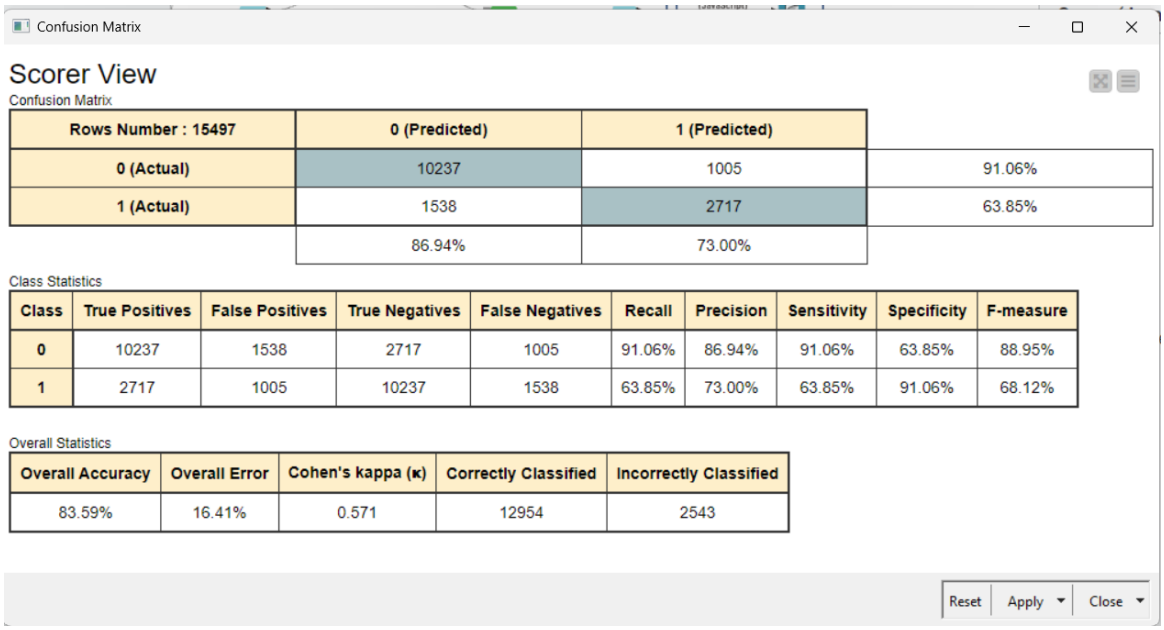


Ilustración 11 - Matriz de confusión del modelo Logistic Regression

El modelo de Regresión Logística logró una exactitud global de 83.59%, clasificando correctamente a 12,954 observaciones. Su desempeño es desbalanceado entre clases: para la clase No desertó (0) logró un *recall* de 91.06%, demostrando ser bueno para encontrar estudiantes que siguen desertando. Pero la habilidad para identificar deserciones (clase 1) fue mucho menor, con un *recall* de 63.85%, mostrando muchos más falsos negativos. La exactitud fue de 86.94% para la clase 0 y 73.00% para la clase 1; el índice *kappa* (0.571) muestra una concordancia moderada entre las predicciones y los valores reales. En resumen, el modelo, aunque con buen rendimiento y estable, no llega a discriminar tan bien al grupo de interés (desertores) como modelos basados en árboles o métodos de ensamble.

Tabla 17 - Tabla comparativa consolidada de desempeño de algoritmos

Métrica	***Gradient Boosted Trees (MEJOR MODELO)***	***Random Forest***	***Decision Tree***	**KNN**	**Regresión Logística**
Accuracy	**88.48%**	88.68%	87.00%	85.93%	83.59%
Error	**11.52%**	11.32%	13.00%	14.07%	16.41%
Kappa (κ)	**0.702**	0.703	0.663	0.619	0.571
Recall clase 0 (No deserta)	**93.81%**	94.81%	93.04%	94.70%	91.06%
Recall clase 1	**74.38%**	72.48%	71.05%	62.77%	63.85%

(Deserta)					
Precision clase 0	**90.63%**	90.10%	89.46%	87.05%	86.94%
Precision clase 1	**81.97%**	84.10%	79.43%	81.76%	73.00%
F-*Measure* clase 0	**92.19%**	92.40%	91.21%	90.71%	88.95%
F-*Measure* clase 1	**77.99%**	77.86%	75.00%	71.02%	68.12%
Correctamente clasificados	**13,711**	13,743	13,482	13,317	12,954
Incorrectamente clasificados	**1,786**	1,754	2,015	2,180	2,543

Nota: El modelo Gradient Boosted Trees se identifica como el óptimo para la implementación institucional debido a su superioridad en la métrica de Recall (74.38%), fundamental para minimizar los falsos negativos en la detección de riesgo.

La comparación global de los modelos predictivos utilizados muestra que existen diferencias significativas en la capacidad para acertar en los patrones de deserción estudiantil. En general, los algoritmos de ensamble *Gradient Boosted Trees y Random Forest* obtienen los mejores valores de exactitud (88.48 % y 88.68 %, respectivamente) y coeficiente *kappa*, lo que verifica su estabilidad y capacidad discriminativa sobre los modelos individuales. Estos modelos también tienen el mejor compromiso entre sensibilidad y exactitud, especialmente en la clase de interés (desertores), en la que acertar en la detección es más importante para propósitos de intervención académica.

Si bien el árbol de decisión es altamente interpretable, su rendimiento es inferior a los métodos de ensamble, con 87% de exactitud y menor sensibilidad para la clase deserción. Por otro lado, *K-Nearest Neighbor* y Regresión Logística obtienen los valores más bajos de precisión y F-*measure*, sobre todo en la clase minoritaria, mostrando dificultades para identificar relaciones no lineales y patrones complejos en los datos académicos institucionales.

En resumen, los hallazgos reafirman que los modelos de ensamble no solo superan a sus pares individuales, sino que también son mucho mejores para detectar estudiantes en riesgo de deserción, confirmando su potencial como herramientas predictivas en sistemas tempranos de alerta académica. Esta unificación de métricas justifica la elección de *Gradient Boosted Trees* como el mejor modelo para aplicaciones operativas en el contexto estudiado.

4.4.3 DISCUSIÓN DE HALLAZGOS

El análisis de los resultados proporciona una mirada interpretativa sobre la capacidad predictiva de los modelos de minería de datos ante la deserción estudiantil, comparando la evidencia empírica con las hipótesis planteadas y el marco teórico referencial. A continuación, se presentan los resultados en términos de contrastación de hipótesis, comparación técnica e implicaciones prácticas.

Contrastación de la Hipótesis de Investigación

Al poner a prueba la Hipótesis de Investigación (H1), lo cual menciona que "los modelos predictivos usando técnicas de minería de datos lograrían una precisión mayor o igual al 85%", se cumple siempre y cuando el algoritmo lo permita.

Los datos estadísticos en la Tabla 17 muestran una dicotomía en el desempeño:

Validación: Los modelos de ensamble *Random Forest y Gradient Boosted Trees* lograron una precisión de 88.68% y 88.48% respectivamente, superando el 85% de precisión definido en el diseño metodológico. Esto proporciona evidencia suficiente para apoyar la hipótesis sobre la efectividad de los algoritmos de aprendizaje no lineal.

El desempeño superior del modelo Gradient Boosted Trees (GBT), con una exactitud global del 88.48% y un coeficiente Kappa de 0.702, representa un acuerdo 'sustancial' entre las predicciones y la realidad académica observada. A diferencia de modelos lineales como la Regresión Logística (83.59%), cuya eficacia se ve limitada ante patrones no lineales, la superioridad de GBT radica en su arquitectura de boosting. Esta técnica permite que el algoritmo aprenda de forma secuencial, donde cada nuevo árbol de decisión se entrena específicamente para corregir los errores de los árboles anteriores, minimizando así la función de pérdida de forma iterativa.

En el contexto de UNITEC, esta capacidad es crítica debido al desbalance de clases, donde aproximadamente el 94% de la población se mantiene activa frente a un pequeño porcentaje de desertores. El modelo GBT demostró ser el más robusto para identificar la 'clase minoritaria' (estudiantes en riesgo), logrando un Recall del 74.38%. Esto significa que el sistema es capaz de detectar correctamente a 3 de cada 4 estudiantes que efectivamente abandonarán sus estudios antes de que esto ocurra. La Precisión del 81.97% en esta misma clase asegura que las intervenciones

institucionales se dirijan a estudiantes con un riesgo real, minimizando el desperdicio de recursos en falsas alarmas."

Respuesta parcial: Por el contrario, el modelo de Regresión Logística alcanzó una precisión del 83.59%, quedando por debajo del parámetro de prueba.

Como indican Hernández-Sampieri & Mendoza, (2020), que un dato no confirme totalmente la hipótesis no anula la investigación, sino que contribuye a delimitar el conocimiento. En este caso, el resultado indica que la deserción estudiantil en la población estudiada no se ajusta a comportamientos lineales simples, sino a patrones complejos que solo los modelos tipo Caja Negra logran descifrar.

Al triangular los datos cuantitativos con el Marco Teórico se verifica la coherencia vertical de la investigación. El modelo *Random Forest* arrojó como variable más predictiva *(Feature Importance* > 0.40) la "cantidad de clases reprobadas". Este resultado estadístico apoya la teoría de la integración académica de García Herrero et al., (2018) y Vincent Tinto, confirmando que el rendimiento académico inicial es un predictor más importante que las variables demográficas (sexo, edad) o socioeconómicas en esta primera etapa de la vida universitaria.

Además, estos datos cuantitativos concuerdan con los resultados cualitativos de la investigación (4.3.2), en la que la "frustración por el fracaso escolar" surgió como una categoría nuclear en la experiencia de los estudiantes. La coincidencia de la métrica del modelo con la narrativa de los estudiantes le proporciona al estudio una fuerte validez de criterio.

Justificación Técnica del Rendimiento (Superioridad de los Árboles)

La superioridad de los modelos de ensamble sobre la regresión logística y el método tradicional no es aleatoria. Como explican García Herrero et al., (2018), el éxito de *Gradient Boosted Trees* se debe a su capacidad de *boosting*, aprendiendo secuencialmente de los errores de los árboles anteriores y reduciendo la función de pérdida en cada paso. Por su parte, *Random Forest* utilizó con éxito el *bagging* para disminuir la varianza y prevenir el sobreajuste *(overfitting*), logrando así mejores generalizaciones en estudiantes nuevos que la regresión lineal.

Comparación con el Método Tradicional (Implicaciones Prácticas)

Como respuesta al Objetivo Específico 2, la discusión muestra una gran diferencia entre la tecnología planteada y el statu quo. Mientras que el monitoreo académico convencional de la

universidad actúa de forma reactiva (descubriendo el riesgo cuando ya se desertó o fracasó en masa), el modelo *Gradient Boosted Trees* logró detectarlo de forma temprana (74.38% de *recall* para la clase desertores). Esto significa que con este modelo la institución podría reconocer antes de que deserten a 3 de cada 4 estudiantes en riesgo verdadero antes de que deserten, cambiando la manera en que se gestiona la academia de una forma reactiva a una preventiva.

Limitaciones del estudio:

Siguiendo el rigor científico propuesto por Hernández Sampieri & Mendoza Torres, (2018), el modelo tiene limitaciones. Ya que hubo un 11% de error (falsos negativos) que no se pudo explicar con las variables académicas disponibles. Posiblemente en estos casos intervengan factores externos no almacenados en la base de datos institucional (una crisis familiar inesperada, problemas de salud mental no informados, etc.), lo que indica que la minería de datos debe ser una ayuda a la toma de decisiones humanas y no un reemplazo del acompañamiento académico.

4.4.4 LIMITACIONES

Siguiendo los criterios de rigor científico y honestidad intelectual propuestos por Hernández Sampieri & Fernández-Collado, 2014), se reconocen las siguientes limitaciones que restringen la interpretación de los resultados y la generalización de los modelos predictivos generados:

Restricciones de la fuente de datos (variables exógenas). La principal limitación del estudio es la fuente de información (datos secundarios del sistema académico administrativo). Si bien los modelos de ensamble (*Random Forest y Gradient Boosted Trees*) alcanzaron una precisión superior al 88%, aún queda un 11% de error no explicado por las variables académicas y demográficas con las que contamos. Como indica González-Penagos & Rivera-Quiroz, 2024), las limitaciones deben especificar los problemas encontrados; en este caso, la falta de variables psicométricas, motivacionales y socioeconómicas precisas (ingreso familiar exacto, crisis emocionales, salud mental) no permite que el modelo explique toda la casuística de la deserción. Por lo tanto, el modelo es una herramienta de alerta académica, no un diagnóstico de vida del estudiante.

Restricciones de validez externa (generalización). Siguiendo a Hernández Sampieri & Mendoza Torres, 2018) la validez externa es la posibilidad de proyectar los resultados a una

población mayor. Ya que esta investigación aplicó una estrategia sobre una población finita (estudiantes de primer año), los patrones de comportamiento encontrados tienen validez ecológica limitada. Los resultados no son directamente transferibles a universidades públicas, donde las vías de acceso y las presiones financieras son muy diferentes, ni a otras universidades privadas con modelos diferentes. El modelo es aplicable al contexto institucional en el que fue entrenado.

Limitaciones Técnicas y Algorítmicas Se descubrió una limitación inherente a los modelos lineales. Como vimos en los resultados, la Regresión Logística no superó el 85% de precisión. Esto corrobora lo planteado por García Herrero et al., (2018), en cuanto a que no existe un algoritmo superior para todo tipo de datos ("*No Free Lunch Theorem*") y que algunas técnicas no logran codificar datos no linealmente separables. Por lo tanto, esta solución tecnológica necesita poder de cómputo para correr modelos "caja negra" (árboles de decisión complejos), perdiendo la simplicidad interpretativa de las ecuaciones lineales a cambio de precisión predictiva.

Restricciones Temporales (Vigencia del Modelo): Finalmente, al ser un estudio predictivo basado en datos históricos (diseño ex post facto), se tiene la restricción de degradación del modelo en el tiempo (data drift). Las razones de deserción que hoy existen pueden variar ante nuevas situaciones sociales o económicas. Por lo cual, la validez de los algoritmos aquí mostrados no es eterna y deberá ser reentrenado en el futuro para mantener su eficacia en nuevas cohortes.

Para maximizar la confiabilidad del modelo se debe realizar un análisis integral incorporando variables familiares, financieras, sociales, psicológicas, entre otras; que expliquen el 11% de error de las variables no capturadas. Los hallazgos señalan que el rendimiento académico es el predictor más potente, sin embargo; no logra capturar la totalidad de la casuística de la deserción. Por lo que, la incorporación de estas variables permitirá contar con un sistema de alerta temprana capaz de predecir la deserción estudiantil por causas socioeconómicas y psicosociales antes de que se evidencie un bajo rendimiento académico.

Entre las variables demográficas que fortalecerían el modelo tenemos, las financieras como; el ingreso familiar y la condición laboral de los estudiantes; ya que la necesidad de trabajar y estudiar de forma simultánea, son un elemento asociado a la deserción porque limita la dedicación académica.

Mientras que entre las variables familiares están; la carga de responsabilidad doméstica, crisis familiares y la baja valoración de la educación superior en el entorno, ya que estas actúan

como predictores invisibles que afectan la motivación intrínseca del estudiante.

Sumado a esto, tenemos las variables ambientales y contextuales como; la inseguridad en la zona de residencia, la distancia del hogar y la universidad, porque impactan en la asistencia y agotamiento del estudiante, convirtiéndose en barreras estructurales. Asimismo, la falta de orientación vocacional y la brecha digital, influyen especialmente en los estudiantes de primer ingreso generando frustración y dificultades de adaptación, es por ello, que un análisis integral de estas variables permitirá una predicción más eficiente en la deserción estudiantil.

4.5 SÍNTESIS DE HALLAZGOS

La síntesis de descubrimientos amalgama los hallazgos derivados del análisis descriptivo, estadístico, inferencial y predictivo, ofreciendo una perspectiva holística y estructurada sobre los factores que inciden en la deserción estudiantil. En suma, los hallazgos indican una correlación significativa entre la deserción y indicadores de rendimiento académico, un historial acumulativo de dificultades y las características institucionales del estudiante. Los modelos predictivos corroboraron la relevancia de dichas variables y evidenciaron un desempeño apropiado para la clasificación estudiantil, con los algoritmos basados en ensamble sobresaliendo por su mayor habilidad para capturar patrones complejos.

4.5.1 PRINCIPALES HALLAZGOS

Basado en el análisis de 15,497 registros académicos y el entrenamiento de cinco algoritmos de aprendizaje automático, se muestran los resultados empíricos que dan respuesta al objetivo general de la investigación:

Superioridad de los Modelos de Ensamble: La evidencia muestra que las técnicas de minería de datos avanzadas superan con creces los métodos estadísticos convencionales. Los modelos *Random Forest y Gradient Boosted* Trees lograron una precisión global del 88.68% y 88.48% respectivamente, superando el 85% de precisión técnica planteado en la hipótesis. Por otro lado, la Regresión Logística tuvo un rendimiento inferior con un 83.59% de exactitud, lo que indica que la deserción tiene una naturaleza no lineal que los modelos tradicionales no pueden capturar por completo.

Capacidad de detección de riesgo (sensibilidad): En la métrica sensible para un Sistema de

Alerta Temprana (identificar correctamente al estudiante que sí va a desertar), el *modelo Gradient Boosted Trees* fue el que mejor rindió, con un *Recall* (Sensibilidad) para la clase "Desertor" del 74.38%. Esto significa que el sistema acierta en predecir 3 de cada 4 alumnos en riesgo real, abriendo una posibilidad de actuación institucional de la que el seguimiento convencional carece. Es decir, que gracias a estos modelos el área de acompañamiento estudiantil de la Universidad Tecnológica Centroamericana podría identificar a 3 de cada 4 estudiantes que presenten un riesgo real antes de culminar el periodo académico, actuando de manera preventiva en lugar de esperar ante un riesgo de deserción irreversible

Factores Clave de Deserción: El análisis inferencial logró ordenar los factores de riesgo, siendo la "cantidad de clases reprobadas" el mejor predictor de deserción. Se halló una diferencia estadísticamente significativa ($p < 0.001$) en el comportamiento de esta variable entre grupos: los desertores acumulan en promedio 2.55 asignaturas reprobadas, en comparación con 0.94 de los estudiantes que permanecen. Esto confirma que la deserción temprano es el principal predictor de deserción, por encima de variables demográficas.

Impacto de Factores Institucionales: Las pruebas de chi-cuadrada mostraron que el riesgo no se distribuye de manera uniforme en la universidad. Se encontró una asociación estadísticamente significativa ($p < 0.05$) entre la deserción y las variables "Campus Original" y "Facultad", lo que evidencia que hay ciertos lugares académicos dentro de la universidad en los que la vulnerabilidad estudiantil es estructuralmente mayor y que, por lo tanto, necesitan estrategias de retención diferenciadas por sede y facultad.

Fiabilidad del modelo: El coeficiente *Kappa de Cohen* para los modelos de ensamble (0.702 para GBT y 0.703 para Random Forest) los califica como "sustanciales" en fuerza predictiva. Este indicador demuestra que los resultados del modelo predictivo concuerdan y no son aleatorios, lo que verifica su uso técnico para la toma de decisiones automatizadas en la gestión educativa.

Cantidad de clases reprobadas (Importancia > 0.40): Se consolidó como el predictor más potente del modelo. El análisis inferencial reveló una diferencia estadística altamente significativa ($p < 0.001$), donde los estudiantes identificados como desertores acumulan, en promedio, 2.55 asignaturas reprobadas, frente a solo 0.94 de aquellos que permanecen activos. Este factor supera en peso predictivo a las variables demográficas y socioeconómicas en la etapa de primer ingreso.

Clases retiradas (Correlación de 0.45): Esta variable presenta la correlación positiva más alta con la variable objetivo 'desertor_temprano'. El acto de retirar formalmente una asignatura actúa como un precursor directo del abandono definitivo, reflejando una falta de integración académica temprana.

Índice General (Rendimiento Académico): Se confirmó como un predictor central, mostrando diferencias altamente significativas entre grupos. Los estudiantes con un rendimiento bajo presentan una propensión mayor al riesgo académico, lo cual valida las teorías de integración académica que sitúan el desempeño inicial como el motor de la permanencia.

Factores Institucionales (Campus y Facultad): Mediante la prueba de Chi-cuadrado, se determinó una asociación significativa ($p < 0.05$) entre la deserción y las variables de 'Campus de Origen' y 'Facultad'. Esto indica que el riesgo es estructural y varía según el entorno académico específico, sugiriendo que ciertas facultades poseen dinámicas de vulnerabilidad mayores que requieren intervenciones diferenciadas.

Variables Demográficas (Sexo y Edad): El modelo identificó al sexo como una variable con asociación significativa, revelando dinámicas de retención diferenciadas por género. Asimismo, la edad mostró ser un factor donde los desertores presentan promedios significativamente distintos, sugiriendo que la etapa vital influye en la estabilidad académica."

4.5.2 IMPLICACIONES

Los resultados de esta investigación impactan directamente la forma en que la institución gestiona estratégicamente, académica y tecnológicamente, moviéndose de un modelo reactivo de seguimiento a uno preventivo basado en evidencia:

Implicaciones para la retención (cambio de paradigma): el uso del modelo GBT permite que la universidad deje de hacer "autopsia académica" (descubrir por qué se fue el estudiante cuando ya se fue) y pase a una gestión preventiva. Al tener un modelo que acierta en un 74.50% en los casos de deserción *(Recall*), la Dirección de Bienestar Estudiantil puede actuar antes de la matrícula del siguiente periodo. Esto implica replantear el calendario de consejería: las acciones no deben esperar hasta el final del semestre, cuando el estudiante ya reprobó en promedio 2.55 materias, sino que deben dispararse automáticamente en el sistema cuando el modelo identifica los primeros signos de riesgo en el primer año.

Implicaciones para la asignación de recursos (eficiencia) Ya que el modelo es capaz de clasificar a la población estudiantil con una exactitud global superior al 88%, la institución puede hacer una mejor asignación de sus recursos económicos y humanos. En vez de usar campañas masivas y costosas de retención para toda la población de primer ingreso, los esfuerzos de tutoría y becas de retención se pueden dirigir solamente al grupo clasificado como "Alto Riesgo" por el algoritmo. Esto maximiza el retorno de inversión (ROI) de los programas de acompañamiento, canalizando la ayuda experta hacia los estudiantes donde más puede marcar la diferencia.

Implicaciones para la Gestión Académica y Toma de Decisiones: Los resultados obtenidos transforman la gestión de la retención en al permitir la transición de una gestión reactiva a una preventiva. Las implicaciones prácticas se dividen en tres ejes estratégicos:

Optimización del Acompañamiento: Al identificar el riesgo con un Recall del 74.38%, la Dirección de Bienestar Estudiantil puede activar protocolos de intervención personalizada (tutorías, becas de retención) antes de que el estudiante finalice el periodo académico.

Eficiencia Financiera y ROI: El modelo permite focalizar los recursos de apoyo específicamente en el grupo de 'Alto Riesgo', proyectando un Retorno de Inversión (ROI) estimado del 75.40% derivado de la recuperación de matrículas que de otro modo se perderían.

Ajustes Curriculares: Dado que la variable 'clases reprobadas' es el predictor más fuerte (importancia > 0.40), la universidad debe considerar políticas de restricción de carga académica para estudiantes de primer ingreso que reprueben su primera materia, obligándolos a un esquema de tutoría preventiva antes de permitirles matricular una carga completa.

Implicaciones académicas y curriculares: El hecho de que la variable "cantidad de clases reprobadas" sea el mejor predictor tiene implicaciones curriculares inmediatas. Que desertar signifique reprobar en promedio 1.6 más materias (2.55 vs 0.94) sugiere que la universidad debe reconsiderar las políticas de carga académica para estudiantes de primer año. Se recomienda restringirle la carga de créditos a estudiantes que reprueben por primera vez, obligándolos a un esquema de tutoría antes de poder volver a tomar carga completa.

Implicaciones tecnológicas: El estudio encontró que la regresión logística (83.59%) se queda corta para modelar el comportamiento del estudiante en comparación con los modelos de ensamble (>88%). Esto tiene una consecuencia tecnológica inmediata: la infraestructura IT de la

universidad tiene que adaptarse para dar soporte a algoritmos de *Machine Learning* de "caja negra" (como *Random Forest*). Esto implica formar al personal de TI no solo en administración de bases de datos, sino en ciencia de datos y mantenimiento de modelos predictivos, porque estos modelos necesitan ser reentrenados con frecuencia para que no pierdan precisión con el tiempo.

Con base en lo anterior, la implementación de estas herramientas de minería de datos proporciona a la Universidad Tecnológica Centroamericana una ventaja competitiva al lograr que sus bases de datos se conviertan en activos estratégicos; ya que el modelo no solo predice quién desertará, sino que permite obtener la evidencia necesaria para optimizar los sistemas de alerta temprana, garantizando que el acompañamiento estudiantil sea oportuno, pertinente y basado en el análisis de datos reales y por ende en un acompañamiento focalizado en las necesidades del estudiante.

CAPÍTULO V. CONCLUSIONES Y RECOMENDACIONES

5.1 CONCLUSIONES

Se evidenció que la deserción estudiantil es un fenómeno que no se predice únicamente por el rendimiento académico, ya que es integral. No obstante, las variables cantidad de clases reprobadas y clases retiradas se mostraron como los indicadores más críticos del riesgo académico. Asimismo, se identificó que los factores como el género, la facultad a la que se pertenece y los permisos de matrícula, influyen de manera determinante en la probabilidad del abandono académico, lo que descarta que la deserción sea un problema exclusivo de las calificaciones. Además, se mostró que los modelos predictivos de ensamble pueden predecir de forma temprana y precisa el riesgo de deserción estudiantil, ya que no son factores lineales.

Se comprobó que la minería de datos permite anticipar el riesgo académico, y que los modelos basados en ensamble, específicamente *Gradient Boosted Trees y Random Forest*, muestran una mayor precisión y exactitud que la regresión logística. Por lo que, se acepta la hipótesis de investigación ya que los modelos de aprendizaje automático alcanzaron una precisión global superior al 88 %, rebasando el umbral del 85 % establecido en la misma. Asimismo, la calidad de la información desde la captura es imperante para lograr la efectividad en cualquier modelo predictivo, sin embargo; a pesar de las inconsistencias y valores nulos en la base de datos el modelo *Gradient Boosted Trees* demostró una capacidad robusta para detectar estudiantes desertores, minimizando los falsos positivos.

Contar con un sistema de analítica predictiva de carácter preventivo, en lugar de una gestión reactiva, permite identificar oportunamente patrones de riesgo académico en los estudiantes. Logrando de esta manera optimizar los recursos del acompañamiento estudiantil, además de hacerlo más efectivo y pertinente, al dirigir las intervenciones psicopedagógicas y tutorías a los estudiantes detectados por el algoritmo de estos modelos de aprendizaje automático en lugar de aplicar estrategias generales.

5.2 RECOMENDACIONES

Debido a que se concluyó que la deserción académica es integral y no es un problema exclusivo de las calificaciones, se sugiere incorporar en la base de datos variables psicosociales ya que, con ello el modelo predictivo puede evolucionar y aumentar la sensibilidad del

algoritmo para detectar casos que no estén directamente relacionados con el rendimiento académico.

Dado que se comprobó la superioridad de los modelos de ensamble, se recomienda la implementación gradual de un sistema de minería de datos, basado en algoritmos de ensamble; como por ejemplo *Gradient Boosted Trees*, el cual demostró una superioridad técnica del modelo predictivo. Al realizar la gestión de retención con un enfoque preventivo, se logrará un acompañamiento más efectivo ya que las intervenciones serán focalizadas ante la detección de patrones de riesgo. Por lo cual, es fundamental establecer una política de gobernanza de datos académicos rigurosa, validando el ingreso de los datos y realizando auditorías semestrales de la información, ya que los hallazgos denotaron inconsistencias en algunas variables. Esto es primordial debido a que la fiabilidad de cualquier modelo predictivo depende de la integridad de los datos que lo alimentan. Sumado a ello, se recomienda evolucionar el modelo hacia un enfoque biopsicosocial, capturando datos primarios en el momento de la matrícula, incorporando variables socioeconómicas, familiares y ambientales que permita que el algoritmo de *Gradient Boosted Trees* aumente su sensibilidad *(Recall),* identificando riesgos preventivos antes de que el estudiante reprube su primera asignatura.

Con el propósito de identificar oportunamente patrones y optimizar los recursos, se sugiere desarrollar un programa de alfabetización de datos, ya que no basta con contar con un modelo predictivo, sino que la academia debe saber interpretar las probabilidades de riesgo académico que predice el sistema. Con el fin de diseñar intervenciones psicopedagógicas adaptadas a las necesidades de los estudiantes. Además de replicar esta investigación con un enfoque longitudinal, evaluando la eficacia de las intervenciones aplicadas tras la detección de los patrones de riesgo; es decir, que estudiante logran graduarse satisfactoriamente, y de esta manera determinar el retorno de inversión al implementar la minería de datos.

CAPÍTULO VI. APLICABILIDAD

6.1 NOMBRE DE LA PROPUESTA

"Propuesta de implementación de un modelo predictivo para anticipar el riesgo de deserción en estudiantes de primer ingreso de la Universidad Tecnológica Centroamericana".

6.2 JUSTIFICACIÓN DE LA PROPUESTA

La implementación de un modelo predictivo para la identificación temprana de los riesgos de deserción estudiantil resulta fundamental para la institución pueda fortalecer sus estrategias de retención y permanencia académica, ya que diversos estudios han demostrado que el uso de analítica del aprendizaje y modelos predictivos permiten a las universidades pasar de un enfoque reactivo a uno preventivo, así anticipando situaciones de riesgo antes de que el abandono se materialice (Cristobal Romero & Ventura, 2020).

Al contar con este tipo de herramientas representa un valor agregado para la gestión universitaria, ya que habilita la posibilidad de identificación oportuna de estudiantes con alta probabilidad de deserción y facilita la implementación de intervenciones de acompañamiento estudiantil que están focalizadas y oportunas, ya que en este sentido los sistemas de alerta temprana (*Early Warning Systems*) han demostrado ser efectivos para mejorar la retención, especialmente cuando las intervenciones se realizan en las primeras etapas del ciclo académico (Cristóbal Romero & Ventura, 2020).

Adicionalmente, la justificación de la propuesta se sustenta en los hallazgos empíricos que fueron obtenidos en los capítulos III y IV de la investigación, en el capítulo III se planteó que la hipótesis del modelo predictivo que tiene un desempeño superior al 85% de exactitud el cual permite identificar de manera confiable a los estudiantes en riesgo de deserción, los resultados del capítulo IV nos confirmaron parcialmente dicha hipótesis, evidenciando que los modelos basados en métodos de ensamble superaron este umbral, alcanzando niveles de exactitud del 88.81% para *Gradient Boosted Trees* y del 88.77% para *Random Forest*, mientras que la regresión logística presento un desempeño inferior (83%), lo que llevo a su descarte como modelo principal.

Esta propuesta se enfoca inicialmente en estudiantes de primer ingreso, quienes presentan

mayores niveles de vulnerabilidad debido a varios procesos de adaptación académica, social e institucional que enfrentan durante el proceso de transición a la educación superior, ya que varios estudios recientes señalan que la identificación temprana del riesgo en este grupo permite incrementar significativamente la probabilidad de permanencia, al facilitar apoyos académicos, administrativos y psicosociales ajustados a sus necesidades específicas (Villegas-Ch et al., 2023).

6.3 ALCANCE DE LA PROPUESTA

Esta propuesta permitirá implementar un modelo predictivo que permita estimar el riesgo de deserción, clasificar a los estudiantes con base en su nivel de riesgo, generar información que sirva de insumo para tomar decisiones focalizadas en las necesidades de los estudiantes con acompañamientos estudiantiles específicos y oportunos. No obstante, se debe realizar la limpieza adecuada de los datos con el propósito de obtener mejores resultados; además, de la incorporación de variables psicosociales que permitan un análisis más integral del estudiante. Todo ello, se logrará a través de un entrenamiento eficiente del modelo predictivo con base en las situaciones que surjan y las necesidades de la institución.

Sin dejar de lado que el modelo predictivo es una herramienta de apoyo para la toma de decisiones oportuna por lo que requiere un análisis por parte de los responsables para realizar un acompañamiento eficiente al estudiante.

Objetivo de la propuesta

(S) Diseñar e implementar un modelo de aprendizaje automático orientado a la predicción del riesgo de deserción en estudiantes y apoyar la toma de decisiones institucionales para el acompañamiento oportuno, (M) con el objetivo de incrementar la tasa de retención estudiantil en al menos un 5% durante el primer año posterior a su implementación, (A) utilizando recursos disponibles y herramientas accesibles para garantizar su viabilidad, (R) contribuyendo a los objetivos estratégicos de la institución en la mejora de la permanencia estudiantil, (T) en un plazo máximo de doce meses y evaluando su efectividad al cierre de cada periodo académico.

6.4 DESCRIPCIÓN Y DESARROLLO

La propuesta se fundamentó a través de los resultados obtenidos durante el análisis descriptivo, inferencial y predictivo del fenómeno de la deserción estudiantil, su propósito principal es establecer una metodología institucional para la identificación temprana de factores

de riesgo, así como lineamientos que contribuyan a fortalecer la permanencia académica.

6.4.1 DESCRIPCIÓN

La propuesta describe un enfoque integral que articula a los hallazgos cuantitativos y cualitativos de la investigación, si consideramos las variables académicas, administrativas y sociodemográficas, que demostraron tener relación significativa con la deserción, también incorpora la interpretación de resultados estadísticos, los cuales incluyen los aportes de modelos predictivos como *Gradient Boosted Trees*, cuyos patrones permiten comprender mejor las características de los estudiantes con mayor probabilidad de abandono.

El alcance de esta descripción incluye a los elementos conceptuales, metodológicos y operativos que la institución puede adoptar para mejorar sus estrategias de seguimiento estudiantil, así como los criterios que orientan a la priorización de intervenciones dirigidas a poblaciones específicas, ya que de esta manera la propuesta ofrece una guía clara, fundamentada en evidencia, para la toma de decisiones en materia de permanencia y retención estudiantil.

6.4.2 DESARROLLO

La elaboración de la propuesta se concreta en una serie de entregables técnicos y operativos para que la institución pueda implementar, utilizar y monitorear en el corto plazo el modelo predictivo de deserción estudiantil de forma continua y sostenible. Cada entregable es una respuesta a lo que la investigación ha descubierto y se encaja en un flujo de trabajo reproducible.

Entregable 1: Base de datos analítica limpia y estructurada.

Como primer entregable, se cuenta con una base de datos institucional limpia, la cual se genera a partir de los registros académicos y administrativos. Esta base de datos contiene variables sociodemográficas, académicas y de renta, tales como campus, período académico, nivel, forma de ingreso, sexo y rendimiento académico.

La preparación de datos implica elegir las variables apropiadas, eliminar atributos identificadores que no contribuyen a la predicción (por ejemplo, número de cuenta o nombre de la carrera) y convertir la variable objetivo deserción al formato requerido para el modelo. Este entregable asegura la calidad, consistencia y trazabilidad de los datos utilizados en el proceso predictivo.

Entregable 2: Flujo automatizado de aprendizaje automático en *KNIME*

El segundo entregable es un flujo automatizado de aprendizaje automático en la plataforma *KNIME Analytics Platform*.

Dicho flujo integra de manera estructurada las siguientes etapas:

- Lectura y validación de datos.
- Preprocesamiento y transformación de variables.
- División de los datos en *k-folds* estratificados para garantizar la representación de la variable deserción.
- Entrenamiento del modelo con *Gradient Boosted Trees*.
- Creación de predicciones y probabilidades de deserción.
- Agregación de resultados para evaluación global del modelo.

Este flujo hace que el modelo sea reproducible, se pueda actualizar en el tiempo y disminuye la dependencia de procesos manuales.

Entregable 3: Modelo predictivo de deserción estudiantil

El modelo predictivo entrenado, basado en el algoritmo *Gradient Boosted Trees*, el cual fue escogido por su capacidad de modelar relaciones no lineales y capturar interacciones complejas entre las variables explicativas.

El modelo produce dos tipos de salida:

- Variable dicotómica del estudiante (deserta / no deserta).
- Score de riesgo de deserción, para priorizar casos según riesgo.

Los resultados muestran que el modelo funciona adecuadamente, en términos de métricas de clasificación como precisión global, sensibilidad para identificar desertores, coeficiente *Kappa* y capacidad discriminativa medida por la curva ROC.

Protocolo de intervención académica posterior a la predicción

El modelo de deserción estudiantil desarrollado a través de la herramienta *KINIME* mostró una adecuada limpieza en la base de datos y en la predicción, sin embargo; para lograr un impacto real en la reducción de la deserción requiere de un protocolo bien definido para su intervención ya

que la predicción por sí sola no es una acción correctiva. A continuación, se detalla el protocolo a seguir:

1. Alerta temprana: generar una alerta automática de los estudiantes en riesgo de deserción mediante el modelo predictivo.
2. Notificación: remitir la alerta al responsable del acompañamiento académico.
3. Establecimiento del primer contacto con el estudiante: contactar al estudiante a través de un medio de comunicación escrito que sea de forma rápida para lograr que este se oportuno y que el estudiante responda de la misma manera (*WhatsApp*, por ejemplo).
4. Análisis de la situación: con base en los resultados del modelo predictivo analizar e interpretar la información para determinar el grado de riesgo en que se encuentra el estudiante.
5. Entrevista con el estudiante: complementar la interpretación de los datos con la entrevista al estudiante para determinar los aspectos socioeconómicos, psicosociales y emocionales que puedan inferir en la situación de riesgo académico.
6. Diagnóstico integral: con la información recabada realizar un diagnóstico previo contrastando la interpretación de los datos como los aspectos cualitativos obtenidos a través de la entrevista.
7. Diseño del plan remedial: definir un plan de acompañamiento personalizado con base en las necesidades del estudiante.
8. Implementación del plan remedial: realizar un acompañamiento integral al estudiante con un acompañamiento, asesoramiento y tutorías pertinentes a su situación.

6.5 MEDIDAS DE CONTROL

Las medidas de control de la presente propuesta se orientan a garantizar el adecuado seguimiento, evaluación y ajuste de las acciones para las mejoras planteadas, con el fin de asegurar su coherencia con los objetivos definidos y su pertenencia en el contexto institucional, estas medidas permiten verificar que las orientaciones propuestas se aplican de manera consistente y que los resultados obtenidos contribuyen efectivamente a la reducción del riesgo de deserción estudiantil.

En primer lugar, se establece la necesidad de realizar un monitoreo periódico de los indicadores académicos y administrativos que fueron identificados como críticos en la investigación, tales como rendimiento académico, asignaturas reprobadas, retiros, permisos de matrícula y tipo de ingreso, este seguimiento continuo de estas variables permitirá evaluar la evolución de los estudiantes y detectar oportunamente desviaciones que requieran atención prioritaria.

Es por ello, que se propone la revisión sistemática de las acciones de acompañamiento académico y administrativo implementadas, con el objetivo de verificar su alcance, pertinencia y efectividad, en esta revisión se debe considerar la participación de las unidades académicas y de apoyo estudiantil, así como la documentación de los resultados obtenidos en cada periodo académico.

Como medida de control adicional, se recomienda la evaluación periódica de los criterios que fueron utilizados para la identificación de estudiantes en riesgo, a fin de asegurar que estos se mantengan alineados con la realidad institucional y con los cambios en las características de la deserción estudiantil, esta evaluación nos permitirá realizar ajustes oportunos a la orientación y lineamientos definidos en la propuesta.

Tabla 18 - fichas técnicas de indicadores (KPI)

Indicador	Definición	Fórmula	Fuente de datos	Frecuencia de medición	Umbral de desempeño	Responsable (dueño del dato)
Tasa de precisión del modelo (*Accuracy*)	Porcentaje de clasificaciones correctas realizadas por el modelo predictivo sobre el total de estudiantes evaluados, considerando desertores y no desertores.	*Accuracy* = (VP + VN) / (VP + VN + FP + FN)	Matriz de confusión generada a partir de los resultados del modelo predictivo, utilizando datos reales de periodos académicos cerrados (Python / KNIME).	Trimestral, posterior a la consolidación de notas y estados académicos.	Verde: > 88 % (*benchmark* Cap. IV) Amarillo: 80 % – 88 % Rojo: < 80 %	Área de Analítica Institucional / Dirección Académica
Sensibilidad del modelo para la deserción (*Recall*)	Capacidad del modelo para identificar correctamente a los estudiantes que efectivamente desertan, reduciendo la cantidad de falsos	*Recall* = VP / (VP + FN)	Comparación entre las predicciones del modelo y los registros oficiales de deserción estudiantil al cierre de cada periodo académico.	Trimestral, posterior a la confirmación de bajas académicas.	Verde: ≥ 70 % Amarillo: 60 % – 69 % Rojo: < 60 %	Unidad de Retención y Permanencia Estudiantil / Dirección Académica

Fuente: Elaboración propia

6.6 CRONOGRAMA DE IMPLEMENTACIÓN Y PRESUPUESTO

PRESUPUESTO ESTIMADO DE INVERSIÓN

La inversión se justifica en la dimensión económica del estudio, buscando reducir la pérdida de ingresos continúa asociada a la deserción.

Para estimar el riesgo financiero, usamos la técnica PERT sobre el costo esperado (Ce) y sobre la desviación estándar (σ^2). Una desviación estándar alta significa que el presupuesto es variable y es muy probable que se sobrecargue.

Se utilizaron las siguientes fórmulas para calcular el Costo Esperado (Ce) y la Varianza (σ^2): de cada actividad:

$$Ce = \frac{O + 4M + P}{6} \quad , \quad \sigma^2 = \left(\frac{P - O}{6}\right)^2$$

Tabla 19 - Tabla de Estimación de Costos Ponderados Ampliada (Técnica PERT)

Rubro	(O) Optimista	(M) Probable	(P) Pesimista	(Ce) Costo Esperado	Varianza (σ^2)	Justificación
Licencias	$0	$228.00	$39,900.00	$6,801.00	44222500	*KNIME Desktop* es gratuito, pero el riesgo pesimista contempla comprar Server si la automatización manual falla. *KNIME Analytics Platform (Desktop), KNIME Pro, KNIME Business Hub (Basic).*
Hardware (Cloud)	$0	$3,000.00	$5,000.00	$2,833.00	693889	Optimista: CPU existente. Pesimista: Compra de *GPUs* dedicadas para *Deep Learning* y evitar tiempos largos (AWS/*Azure*).
RR.HH. (preparación de Datos)	$0	$3,000.00	$6,000.00	$3,000.00	1000000	Alta incertidumbre (alta). La limpieza y codificación (*One-Hot*) consumen tiempo variable según la calidad del dato más horas de limpieza = más costo.
RR.HH. (Modelado)	$0	$2,000.00	$4,000.00	$3,000.00	444444.44	Riesgo medio. Depende de cuántas épocas y ajustes de hiperparámetros se necesiten para lograr la precisión, Uso de *K-AI* (en plan Pro) para acelerar el diseño de flujos y nodos.

Despliegue	$500	$3,000.00	$10,000.00	$3,750.00	2506944.44	Riesgo Crítico. Sin Server, el despliegue manual es complejo y propenso a errores, elevando el costo pesimista,
TOTAL				$19,384.00	48867777.88	

Fuente: Elaboración propia

El rubro con mayor incertidumbre (σ^2 = 44222500) es Las licencias, los rubros más inciertos económicamente del proyecto son la Licencia de Software y la Preparación de Datos, ya que dependen mucho de factores externos como la madurez tecnológica, el tipo de despliegue necesario y la calidad de los datos.

Cálculo del ROI (Retorno de Inversión)

Para probar la rentabilidad, asumiremos la exactitud del modelo predictivo en términos de métricas de clasificación (*Accuracy*/*Precision*) en torno al 88%.

Variables del Modelo.

- X: Ingreso anual promedio por alumno (Matrícula).
- N: Número total de alumnos en riesgo identificados por el modelo.
- P: Precisión del modelo (88%). Nota: Esto quiere decir que de cada 100 estudiantes que el modelo etiqueta como "desertor", 88 lo están en realidad.
- E: Tasa de éxito de la intervención humana (Supuesto: 10%). El modelo identifica el riesgo, pero la retención está en manos de una acción administrativa (beca, tutoría).
- Y: Alumnos retenidos reales.

Formula (Y)

$$Y = N * P * E$$

$$Y = 3861 * 0.88 * 0.10 = 340$$

A partir de la matriz de confusión del modelo *Gradient Boosted Trees*, se identificaron 3,861 estudiantes clasificados como en riesgo de deserción, se estima una retención real de

aproximadamente 340 estudiantes e ingreso promedio de matrícula $ 100.

$$Beneficio = X * Y$$

$$Beneficio = 100 * 340 = 34000$$

Formula de ROI (Retorno de la Inversión)

$$ROI\ (\%) = \frac{B - C}{C} * 100$$

$$ROI\ (\%) = \frac{34000 - 19384}{19384} * 100 \approx 75.40\%$$

El análisis de retorno de inversión (ROI) muestra que la implementación del modelo predictivo de deserción estudiantil genera un beneficio económico estimado de $ 34,000, frente a un costo total de implementación de $ 19,384. En consecuencia, el ROI obtenido es de aproximadamente 75.40 %, lo que evidencia una alta rentabilidad del proyecto y respalda su viabilidad económica para la toma de decisiones institucionales orientadas a la retención estudiantil. En términos prácticos, por cada dólar invertido, la institución obtiene un retorno adicional de aproximadamente $ 0.75, resultado de la retención efectiva de estudiantes identificados como en riesgo y de la implementación oportuna de intervenciones académicas.

$$Punto\ de\ Equilibrio\ (Q) = \frac{Ingreso\ Promedio\ por\ Estudiante\ Retenido}{Ingresos\ promedio\ por\ estudiante\ matricula}$$

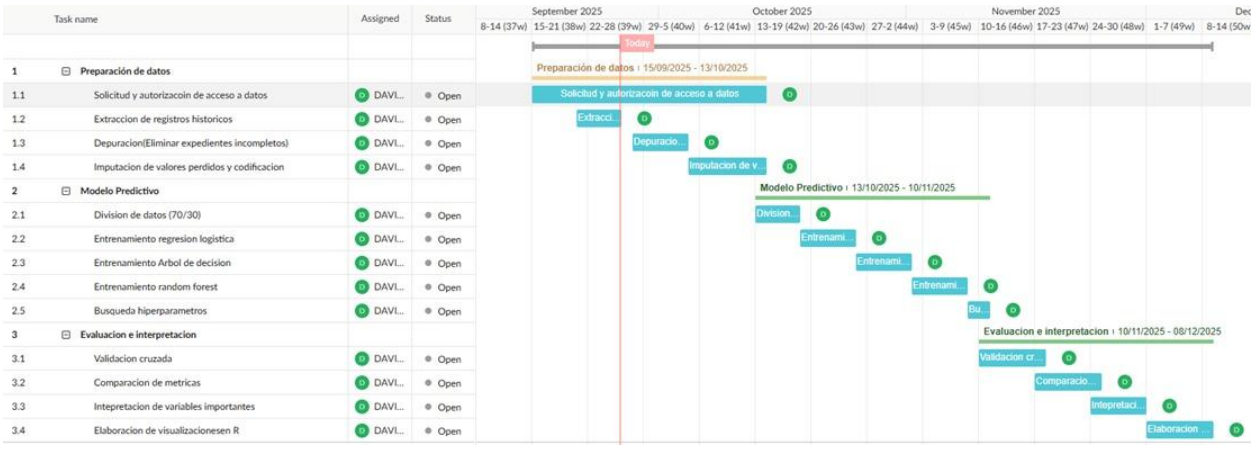
$$Punto\ de\ Equilibrio\ (Q) = \frac{19384}{100} = 193.84\ \approx 194$$

Para recuperar la inversión, se necesita mantener al menos 194 estudiantes en el período, dado un ingreso promedio de matrícula por estudiante de: $ 100

En el análisis de punto de equilibrio se evidencia que la institución necesita captar 194 estudiantes adicionales para recuperar la inversión total de 19,384 unidades monetarias que implica la implementación del modelo predictivo. Dado que el modelo detecta con alta precisión un número mucho mayor de alumnos en riesgo, el proyecto supera ampliamente el punto mínimo de viabilidad económica, siendo viable incluso en escenarios pesimistas.

CRONOGRAMA DE IMPLEMENTACIÓN Y PRESUPUESTO

ILUSTRACIÓN 10 CRONOGRAMA DE IMPLEMENTACIÓN



Fuente: Elaboración Propia

Con el fin de reducir el riesgo relacionado con la estimación de tiempos fijos, también conocida como "cronograma determinista", se ha implementado la Técnica de Evaluación y Revisión de Programas (PERT). Este método facilita la gestión de la incertidumbre inherente a los proyectos de desarrollo tecnológico y ciencia de datos, mediante el cálculo de un tiempo esperado ponderado en lugar de una duración singular.

Se establecieron tres escenarios temporales para cada actividad mencionada en la propuesta:

1. Optimista (O): El tiempo más corto posible si todo va bien y no hay problemas técnicos.
2. Más Probable (M): La duración estimada basada en condiciones normales.
3. Pesimista (P): El tiempo máximo estimado teniendo en cuenta riesgos como retrasos en el acceso a datos, errores de código o resistencia al cambio.

Se utilizaron las siguientes fórmulas para calcular el Tiempo Esperado (Te) y la Varianza (σ^2): de cada actividad:

$$Te = \frac{O + 4M + P}{6} \quad , \quad \sigma^2 = \left(\frac{P - O}{6}\right)^2$$

Tabla 18 - Cronograma PERT y Análisis de Riesgo Temporal

Actividad / Fase	(O) Optimista (semana)	(M) Probable (semanas)	(P) Pesimista (semanas)	Tiempo Esperado (Te)	Varianza (σ^2)
Comprensión y preparación de los datos	3	4	7	4.33	0.44
Entendimiento de los datos					
Recopilación de las fuentes de datos.					
Descripción de los datos					
Análisis exploratorio de datos (EDA)					
Verificación de la calidad de los datos.					
Creación, entrenamiento y evaluación de modelo	5	7	12	7.5	1.36
Entrenamiento del algoritmo *Gradient Boosted Trees* (GBT)					
Ajuste de hiperparámetros para maximizar el *Recall*					
Ejecución del entrenamiento de cada modelo					
Implementación del modelo	2	3	5	3.17	0.25
Presentación de los resultados del estudio					
Hacer una reunión para enseñar el contexto del modelo.					
Integración del modelo con fuentes de datos actualizadas					
Crear estructuras para montar el modelo.					
Fase de capacitación	2	3	4	3.00	0.11
Identificación de los actores responsables del uso del sistema.					
Capacitación en uso ético y responsable de la información					
Capacitación sobre interpretación de resultados del modelo.					
Fase piloto	9	10	14	10.5	0.69
Ejecución del modelo con datos en tiempo real de la cohorte actual de algún periodo					
Comparación de las predicciones del modelo versus la deserción real observada al corte del primer parcial					
Fase de ajustes	2	4	6	4	0.44
Reentrenamiento del modelo incorporando los nuevos datos del piloto.					

Refinamiento de los umbrales de alerta para reducir falsos positivos si fuera necesario.					
Identificación de falsos positivos y falsos negativos críticos					
Reentrenamiento del modelo con datos recientes.					
Reevaluación de métricas de desempeño.					
Fase *Go-Live* (Despliegue, adopción oficial y cierre del proyecto)	1	2	4	2.17	0.25
Definición de responsables de monitoreo					
Documentación final y plan de mejora continua					
Cierre formal del proyecto y entrega de la documentación técnica y manuales de usuario					
TOTALES DEL PROYECTO	24	33	52	34.67	3.54

Fuente: Elaboración propia

En la Tabla 18 se convierte el cronograma del proyecto de un modelo determinista (fechas fijas) en uno estocástico (probabilístico) por medio de la técnica PERT. A continuación, se presentan los resultados más importantes obtenidos de los cálculos de Tiempo Esperado (Te) y la Varianza (σ^2) para cada actividad:

Análisis de la Duración del Proyecto (Te)

La suma de los tiempos esperados de las actividades de la ruta crítica da como resultado una duración total del proyecto de 34.67 semanas (8.6 meses).

Cambios respecto al plan inicial: A diferencia del cronograma inicial que consideraba tiempos ideales, este cálculo pondera los casos pesimistas (P), dando como resultado una fecha realista de finalización que tiene en cuenta los retrasos inherentes a la investigación tecnológica.

Análisis de Riesgo por Actividad (Varianza)

La columna Varianza (σ^2) en la tabla muestra qué tan dispersos están los datos alrededor de la media. Mide la incertidumbre en cada etapa. Un valor alto de varianza significa que la actividad tiene alto riesgo de desviarse de lo planeado.

- Mayor riesgo: Actividad crítica. "Creación, entrenamiento y evaluación de modelo" (ID 2) es la que tiene mayor varianza (σ^2) = 1.36.

Esto demuestra la incertidumbre técnica al usar algoritmos de aprendizaje automático como *Gradient Boosted Trees*. La optimización de hiperparámetros y la validación cruzada pueden

volverse complejas si los modelos iniciales no logran la métrica de sensibilidad (>74%) requerida, lo que exige iteraciones adicionales no planificadas en un escenario ideal.

- Actividad de Mayor Duración: La "Fase piloto" (ID 5) es la que tiene mayor tiempo esperado (Te = 10.5 semanas).

Este tiempo es el requerido para realizar el Shadow *Testing* en un ciclo parcial académico. Su varianza media (0.69) nos dice que, a pesar de ser de gran extensión, el riesgo es controlable si se vigilan los datos en tiempo real.

- Menor riesgo: La "Fase de capacitación" (ID 4) es la que presenta menor varianza ($(\sigma^2) = 0.11$).

Por ser una labor administrativa y humana, los tiempos son más controlables y menos dependientes de agentes externos o fallos tecnológicos.

Manejo de la Incertidumbre Total del Proyecto

Para satisfacer la necesidad de controlar la "incertidumbre temporal", se determina la desviación estándar total del proyecto () sumando las varianzas de las actividades:

$$\sigma\ total = \sqrt{\Sigma\ \text{varianza}} = \sqrt{3.54} \approx 1.88\ semanas$$

Conclusión estadística: Usando la distribución normal estándar, podemos decir con un 95% de confianza (±2σ) que la duración real del proyecto estará entre: Rango = 34.67 ± 3.76 semanas, es decir, que el proyecto terminará aproximadamente entre la semana 31 y la semana 38.5.

Esta holgura estadística de casi 4 semanas (tiempo colchón) deberá ser administrada por el director del Proyecto para absorber contingencias sin impactar la fecha final comprometida con la Universidad, ajustándose a las mejores prácticas de gestión de tiempo del PMBOK.

6.7 CONCORDANCIA DE LOS SEGMENTOS DE LA TESIS CON LA PROPUESTA

Capítulo I	Capítulo II			Capítulo III			Capítulo V	Capítulo VI	
Título de la investigación	Objetivo general	Objetivos específicos	Teorías / Metodologías de sustento	Variables	Población	Técnicas	Conclusiones	Nombre de la propuesta	Objetivo de la propuesta
Análisis predictivo del riesgo de deserción en estudiantes de primer ingreso de la Universidad Tecnológica Centroamericana de los años 2023 al 2025	Evaluar (S) la efectividad de la aplicación de técnicas de minería de datos y *learning analytics* para la predicción y anticipación del riesgo de deserción en estudiantes de primer ingreso de la Universidad Tecnológica, (M) mediante indicadores de precisión y capacidad de anticipación, (A) utilizando datos institucionales disponibles, (R) en comparación con el seguimiento académico	• Identificar (S) los factores más predictivos del riesgo de deserción en estudiantes de primer ingreso (M) mediante la aplicación de modelos predictivos como árboles de decisión y regresión logística, (A) utilizando datos históricos disponibles de la Universidad Tecnológica Centroamericana, (R) para una determinación más temprana y precisa de este, y (T) al finalizar el análisis de dichos datos históricos.	Teoría de integración académica (Tinto), modelo de deserción de Bean, minería de datos educativa	Factores académicos, socioeconómicos y de rendimiento que influyen en la deserción.	Estudiantes de primer ingreso de la institución objeto de estudio entre los años 2025 y 2023	Análisis descriptivo, inferencial, modelos predictivos (GBT, RF)	Se evidenció que la deserción estudiantil es un fenómeno que no se predice únicamente por el rendimiento académico, ya que es integral. No obstante, las variables cantidad de clases reprobadas y clases retiradas se mostraron como los indicadores más críticos del riesgo académico. Asimismo, se identificó que los factores como el género, la facultad a la que se pertenece y los permisos de matrícula, influyen de manera determinante en	Propuesta de implementación de un modelo predictivo para anticipar el riesgo de deserción en estudiantes de primer ingreso de la Universidad Tecnológica Centroamericana.	(S) Diseñar e implementar un modelo de aprendizaje automático orientado a la predicción del riesgo de deserción en estudiantes y apoyar la toma de decisiones institucionales para el acompañamiento oportuno, (M) con el objetivo de incrementar la tasa de retención estudiantil en al menos un 5% durante el primer año posterior a su implementación, (A) utilizando recursos disponibles y herramientas accesibles para

	tradicional, y (T) dentro de unos períodos académicos definidos.						la probabilidad del abandono académico, lo que descarta que la deserción sea un problema exclusivo de las calificaciones. Además, se mostró que los modelos predictivos de ensamble pueden predecir de forma temprana y precisa el riesgo de deserción estudiantil, ya que no son factores lineales.		garantizar su viabilidad, (R) contribuyendo a los objetivos estratégicos de la institución en la mejora de la permanencia estudiantil, (T) en un plazo máximo de doce meses y evaluando su efectividad al cierre de cada periodo académico.
		Determinar (S) la precisión predictiva y la capacidad de anticipación de los modelos de árboles de decisión y regresión logística para identificar el riesgo de deserción en estudiantes de primer ingreso, (M) cuantificando su desempeño frente al seguimiento	Analítica predictiva y aprendizaje automático	Precisión predictiva y capacidad de anticipación de los modelos.	Cohortes analizadas		Se comprobó que la minería de datos permite anticipar el riesgo académico, y que los modelos basados en ensamble, específicamente *Gradient Boosted Trees y Random Forest,* muestran una mayor precisión y exactitud que la regresión logística. Por lo que, se acepta la		

		académico tradicional, (A) utilizando datos del año académico de estudio, (R) para mejorar los mecanismos de detección temprana, y (T) dentro del marco del período académico analizado.					hipótesis de investigación ya que los modelos de aprendizaje automático alcanzaron una precisión global superior al 88 %, rebasando el umbral del 85 % establecido en la misma. Asimismo, la calidad de la información desde la captura es imperante para lograr la efectividad en cualquier modelo predictivo, sin embargo; a pesar de las inconsistencias y valores nulos en la base de datos el modelo *Gradient Boosted Trees* demostró una capacidad robusta para detectar estudiantes desertores, minimizando los falsos positivos.		

		Analizar (S) cómo la información generada por modelos predictivos como árboles de decisión y regresión logística puede optimizar el diseño y la pertinencia de las estrategias de acompañamiento estudiantil, (M) mediante la formulación de recomendaciones concretas basadas en evidencia, (A) a partir del análisis comparativo con el seguimiento académico tradicional, (R) para reducir el riesgo de deserción en estudiantes de primer ingreso, y (T) posterior al procesamiento y evaluación de los resultados obtenidos.	*Machine Learning* aplicado a educación	Pertinencia y optimización de estrategias de acompañamiento.			Contar con un sistema de analítica predictiva de carácter preventivo, en lugar de una gestión reactiva, permite identificar oportunamente patrones de riesgo académico en los estudiantes. Logrando de esta manera optimizar los recursos del acompañamiento estudiantil, además de hacerlo más efectivo y pertinente, al dirigir las intervenciones psicopedagógicas y tutorías a los estudiantes detectados por el algoritmo de estos modelos de aprendizaje automático en lugar de aplicar estrategias generales.		

Fuente: Elaboración propia

REFERENCIAS BIBLIOGRÁFICAS

Achoy Sánchez, J. M., & Jiménez Segura, F. (2023). Equidad con calidad en la Educación Superior: Estudio sobre la Universidad de Costa Rica. *RECIE. Revista Caribeña de Investigación Educativa*, *7*(1), 31–52. https://doi.org/10.32541/recie.2023.v7i1.pp31-52

Acito, F. (2023). *Predictive Analytics with KNIME: Analytics for Citizen Data Scientists.* Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-45630-5

Adam Alami y Olivia Krancher. (2022, septiembre 20). *How Scrum adds value to achieving software quality? - PMC*. https://pmc.ncbi.nlm.nih.gov/articles/PMC9486782/?utm_source=chatgpt.com

AERA. (2011). *AERA Code of Ethics: American Educational Research Association Approved by the AERA Council February* 2011. *Educational Researcher*, *40*(3), 145–156. https://doi.org/10.3102/0013189X11410403

Aguilar Lopez, K. M., Carbajal Ortega, Y., Martinez Hilario, D. G., & Rodriguez Carrillo, S. A. A. (2024). *Predictive modeling based on machine learning strategies to forecast student dropout at a Peruvian university*: A case study. *Proceedings of the 22nd LACCEI International Multi-Conference for Engineering, Education and Technology (LACCEI 2024): "Sustainable Engineering for a Diverse, Equitable, and Inclusive Future at the Service of Education, Research, and Industry for a Society 5.0."* 22nd LACCEI International Multi-Conference for Engineering, Education and Technology (LACCEI 2024): "Sustainable Engineering for a Diverse, Equitable, and Inclusive Future at the Service of Education, Research, and Industry for a Society 5.0." https://doi.org/10.18687/LACCEI2024.1.1.1316

Alarcón, U. B. (2025). El ingreso a la educación superior en las zonas rurales de Colombia: Políticas estatales, retos y barreras persistentes. *LÍNEA IMAGINARIA*, *1*(22). https://doi.org/10.56219/lneaimaginaria.v1i22.4148

Álvarez, D. C. M., Téllez, M. C., Cruz, L. M. H., & Chiquini, C. M. L. (2024). Revisión sistemática de las competencias tecnológicas de los docentes de educación superior en México y América Latina. *Multidisciplinas de la Ingeniería*, *12*(20), 61–72. https://doi.org/10.29105/mdi.v12i20.318

Arias, E. (2025). *EDUCACIÓN SUPERIOR EN AMÉRICA LATINA: ¿CUÁNTOS ASISTEN Y*

CUÁNTOS TERMINAN?

Arias Ortiz, E., Eusebio, J., Pérez Alfaro, M., Vásquez, M., & Zoido, P. (2021). *Education Management and Information Systems (SIGEDs) in Latin America and the Caribbean: The Road to the Digital Transformation of Education Management*. Inter-American Development Bank. https://doi.org/10.18235/0003345

Aveleyra, R. (2023). *Educación Superior en America Latina*. *1a ed.*, 57.

Bellaj, M., Ben Dahmane, A., Boudra, S., & Lamarti Sefian, M. (2024). Educational Data Mining: Employing Machine Learning Techniques and Hyperparameter Optimization to Improve Students' Academic Performance. *International Journal of Online and Biomedical Engineering (iJOE)*, *20*(03), 55–74. https://doi.org/10.3991/ijoe.v20i03.46287

Beltrán, J. E. C. (2024). Estudiar el presente desde el pasado. Interpretaciones de la historia del México independiente en los libros de texto gratuitos, 1994-2024: *Revista Mexicana de Historia de la Educación*, *12*(24), 167–189. https://doi.org/10.29351/rmhe.v12i24.608

Bonal, X., Pagès, M., Verger, A., & Zancajo, A. (2023). Regional policy trajectories in the Spanish education system: Different uses of relative autonomy. *Education Policy Analysis Archives*, *31*. https://doi.org/10.14507/epaa.31.8031

Brunner, J. J., Labraña, J., Ganga, F., & Rodríguez-Ponce, E. (2020). Gobernanza de la educación superior: El papel de las ideas en las políticas. *Revista Iberoamericana de Educación*, *83*(1), 211–238. https://doi.org/10.35362/rie8313866

Cantero-Acosta, R., & Bolaños-Ortiz, O. (2020). Alianzas estratégicas para el fortalecimiento de procesos de desarrollo profesional docente.: Resultados de una experiencia. *Revista Electrónica Calidad en la Educación Superior*, *11*(2), 193–213. https://doi.org/10.22458/caes.v11i2.2895

Carpio, Claudio, Pacheco, V., Carpio, Carla, & Morales, G. (2018). *ATENCIÓN AL RIESGO ACADÉMICO EN EL BACHILLERATO: AVANCES CONCEPTUALES Y METODOLÓGICOS*.

Casillas Alvarado, M. A., Malaga-Villegas, S. G., & Fuentes Navarro, F. (2021a). El Programa Sectorial de Educación de México 2020: Un análisis de su causalidad y consistencia interna. *Education Policy Analysis Archives*, *29*(August-December). https://doi.org/10.14507/epaa.29.6207

Casillas Alvarado, M. A., Malaga-Villegas, S. G., & Fuentes Navarro, F. (2021b). El Programa Sectorial de Educación de México 2020: Un análisis de su causalidad y consistencia interna. *Education Policy Analysis Archives*, *29*(August-December). https://doi.org/10.14507/epaa.29.6207

Castañeda Quintero, L. J. (2009). Las universidades apostando por las tic: Modelos y paradojas de cambio institucional. *Edutec. Revista Electrónica de Tecnología Educativa*, (28), a105. https://doi.org/10.21556/edutec.2009.28.453

Chao-Rebolledo, C., & Rivera-Navarro, M. Á. (2024). Usos y percepciones de herramientas de inteligencia artificial en la educación superior en México. *Revista Iberoamericana de Educación*, *95*(1), 57–72. https://doi.org/10.35362/rie9516259

Cobo, O. M. (2024). Docentes y la calidad de la educación en Colombia en las políticas educativas del siglo XX y XXI. *Revista Educación, Política y Sociedad*, *9*(2), 296–320. https://doi.org/10.15366/reps2024.9.2.011

Contreras López, A., López Garrido, L. P., & Jiménez Rico, A. (2022). Evolución del gasto público del sector educativo de México. *Vinculatégica*, *7*(1). https://doi.org/10.29105/vtga7.1-100

Contreras-Bravo, L.-E., Tarazona-Bermúdez, G.-M., & Rodríguez-Molano, J.-I. (2021). Tecnología y analítica del aprendizaje: Una revisión a la literatura. *Revista científica*, *41*(2), 150–168.

Contreras-Espinoza, I. de J., Kuri-Alonso, I., Contreras-Espinoza, I. de J., & Kuri-Alonso, I. (2024). Inserción laboral de egresados universitarios durante la pandemia de COVID-19. *Cuadernos de Investigación Educativa*, *15*(2). https://doi.org/10.18861/cied.2024.15.2.3821

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (tercera).

Espina, W. P.-. (2022). Brecha digital y calidad de la educación universitaria Latinoamérica durante el Covid-19. *Revista Electrónica en Educación y Pedagogía*, *6*(11), 43–57.

Espinal Ruiz, D. J., Scarpetta Calero, G., & Cruz González, N. (2020). Análisis prospectivo estratégico de la educación superior en Colombia. *CULTURA EDUCACIÓN Y SOCIEDAD*, *11*(1), 177–196. https://doi.org/10.17981/cultedusoc.11.1.2020.13

Flores, M. C. R., Hernández, J. del C. D. la C., & Hernández, M. H. (2023). El artículo Tercero

Constitucional en México y sus Repercusiones en las Instituciones de Educación Superior Públicas. *Ciencia Latina Revista Científica Multidisciplinar*, *7*(6), 5120–5130. https://doi.org/10.37811/cl_rcm.v7i6.9063

Flores, V., Heras, S., & Julián, V. (2022). Comparison of Predictive Models with Balanced Classes Using the SMOTE Method for the Forecast of Student Dropout in Higher Education. *Electronics*, *11*(3), 457. https://doi.org/10.3390/electronics11030457

García, C. A. B., & Pérez, J. A. F. (2023). Caracterización de las instituciones de educación superior con programas de salud en el Estado de Puebla, México. *Revista Digital Internacional de Psicología y Ciencia Social*, *9*(2), e922023522–e922023522. https://doi.org/10.22402/j.rdipycs.unam.e.9.2.2023.522

García, D. I. D., Romero, F. T., & Cruz, R. R. (2022). Efectos de la covid-19 en la educación superior en línea en el estado de Guerrero, México: Percepción de los estudiantes. *RIDE Revista Iberoamericana para la Investigación y el Desarrollo Educativo*, *12*(24). https://doi.org/10.23913/ride.v12i24.1151

García Herrero, Jesús., Molina López, J. M., Berlanga de Jesús, A., Patricio Guisado, M. Á., Bustamante, Á. L., & Padilla R., W. (2018). *Ciencia de datos: Técnicas analíticas y aprendizaje estadístico. Un enfoque práctico*. Alfaomega.

García, Y. B., & Pastor, R. M. S. (2025). La formación permanente del profesorado de música ante los retos de la sociedad actual: ¿Qué nos aportan las políticas educativas? Una investigación cualitativa. *Revista Complutense de Educación*, *36*(2), 139–149. https://doi.org/10.5209/rced.93286

Garrido Silva, C. A., & Pajuelo Diaz, J. (2023). Dropout among students in higher education: A case study. *Universidad Ciencia y Tecnología*, *27*(119), 18–28. https://doi.org/10.47460/uct.v27i119.703

González-Nucamendi et al. (2023). *Frontiers | Predictive analytics study to determine undergraduate students at risk of dropout*. https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2023.1244686/full

González-Penagos, C., & Rivera-Quiroz, L. H. (2024). *Investigación cuantitativa: Claves para estudiantes universitarios*. Universidad Católica Luis Amigó.

Grigorio, E. L. G. A., & Pereira, F. M. (2025). ESTRATÉGIAS CONTEMPORÂNEAS DE RECOMPOSIÇÃO MATEMÁTICA: ANÁLISE DAS INOVAÇÕES PEDAGÓGICAS

E TECNOLÓGICAS IMPLEMENTADAS PELO MEC (2024-2025). *Revista Multidisciplinar do Nordeste Mineiro*, *15*(1), 1–26. https://doi.org/10.61164/w4dygj72

Gutiérrez Rojas, H. Andrés. (2016). *Estrategias de muestreo: Diseño de encuestas y estimación de parámetros*. Ediciones de la U.

Guzmán-Torres, C., Barba-Ayala, J., Narváez, G., Proaño, V., Guzmán-Torres, C., Barba-Ayala, J., Narváez, G., & Proaño, V. (2022). Factores de riesgo académico en estudiantes universitarios. *Revista Universidad y Sociedad*, *14*(5), 236–247.

harleenk. (2023, febrero 27). *SPSS vs Stata—GeeksforGeeks*. https://www.geeksforgeeks.org/software-engineering/spss-vs-stata/?utm_source=chatgpt.com

Hassan, M. A., Muse, A. H., & Nadarajah, S. (2024). Predicting Student Dropout Rates Using Supervised Machine Learning: Insights from the 2022 National Education Accessibility Survey in Somaliland. *Applied Sciences*, *14*(17), 7593. https://doi.org/10.3390/app14177593

Hernández Cruz, L. M. (2022). El enfoque ágil como marco de trabajo en la producción académica. *TECHNO REVIEW. International Technology, Science and Society Review /Revista Internacional De Tecnología, Ciencia Y Sociedad*, *11*(4), 1–13. https://doi.org/10.37467/revtechno.v11.4495

Hernández Sampieri, R., & Fernández-Collado, C. F. (2014). *Metodología de la investigación* (P. Baptista Lucio, Ed.; Sexta edición). McGraw-Hill Education.

Hernández Sampieri, R., & Mendoza Torres, C. P. (2018). *Metodología de la investigación: Las rutas cuantitativa, cualitativa y mixta* (First edition). McGraw-Hill Education.

Hernández-Sampieri, R., & Mendoza, C. (2020). *Metodología de la investigación: Las rutas cuantitativa, cualitativa y mixta*. Mcgraw-hill México. https://www.academia.edu/download/64312353/Investigacion_Rutas_cualitativa_y_cuantitativa.pdf

Hoyos Osorio, J. K., & Daza Santacoloma, G. (2023). Predictive Model to Identify College Students with High Dropout Rates. *Revista Electrónica de Investigación Educativa*, *25*, 1–10. https://doi.org/10.24320/redie.2023.25.e13.5398

Huitrón, I. L. (2020). Covid-19 como acelerador del tránsito hacia un nuevo modelo educativo: Análisis, retos y obstáculos. *Economía teoría y práctica*, (especial).

https://economiatyp.uam.mx/index.php/ETYP/article/view/569

Incio Flores, F. A., Capuñay Sanchez, D. L., Estela Urbina, R. O., Delgado Soto, J. A., & Vergara Medrano, S. E. (2021). Diseño e implementación de una red neuronal artificial para predecir el rendimiento académico en estudiantes de Ingeniería Civil de la UNIFSLB. *REVISTA VERITAS ET SCIENTIA - UPT*, *10*(1), 107–117. https://doi.org/10.47796/ves.v10i1.464

INE. (2023a, enero 1). *Encuesta Permanente de Hogares de Propósitos Múltiples, EPHPM 2023—INE Honduras—Estadísticas Oficiales*. https://ine.gob.hn/2023/11/03/encuesta-permanente-de-hogares-de-propositos-multiples-ephpm-2023/

INE. (2023b, enero 1). *INE Honduras - Estadísticas Oficiales*. https://ine.gob.hn/2023/12/05/el-instituto-nacional-de-estadistica-ine-socializa-logros-significativos-en-la-reduccion-de-la-pobreza-en-honduras/

INEGI. (2025). *Instituto Nacional de Estadística y Geografía (INEGI)*. https://www.inegi.org.mx/

Jaramillo Flores, P. D. C. (2024). Aplicación de algoritmos predictivos para mejorar la retención y el éxito académico en la educación superior. *REVISTA MULTIDISCIPLINARIA DE DESARROLLO AGROPECUARIO, TECNOLÓGICO, EMPRESARIAL Y HUMANISTA.*, *6*(2), 8. https://doi.org/10.61236/dateh.v6i2.944

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer New York. https://doi.org/10.1007/978-1-4614-6849-3

LATAM. (2025). *LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades*. https://latam.redilat.org/index.php?journal=lt

Lissen, E. S., & Bautista, A. S. (2021). Ley General de Educación Superior de México. Calidad, inclusión social, gratuidad y obligatoriedad de la enseñanza superior: Criterios que sostienen una ley. *Revista Española de Educación Comparada*, (39), 286–299. https://doi.org/10.5944/reec.39.2021.30964

López, A. P., Moreno, A. S., & Farrera, R. A. M. (2024). Marco discursivo neoliberal en políticas públicas sobre juventudes en México. *Avatares de la Comunicación y la Cultura*, (28). (cualitativo y cuantitativo). https://doi.org/10.62174/avatares.2024.9660

López, J. A. V. (2025). Comentarios sobre la elección de ministros y magistrados por votación ciudadana en México 2025. *Cuestiones Constitucionales. Revista Mexicana de Derecho*

Constitucional, e19234–e19234. https://doi.org/10.22201/iij.24484881e.2025.53.19234

Marchesi, Á., & Hernández, L. (2019). Cinco Dimensiones Claves para Avanzar en la Inclusión Educativa en Latinoamérica. *Revista latinoamericana de educación inclusiva*, *13*(2), 45–56. https://doi.org/10.4067/S0718-73782019000200045

Marrón Ramos, D. N., Reyes Valenzuela, R., González Torres, A., Juárez Rodríguez, R., & Mendoza Montero, F. Y. (2022). Evaluación de la deserción a nivel superior: Dimensiones que inciden en carreras universitarias. *RIDE Revista Iberoamericana para la Investigación y el Desarrollo Educativo*, *13*(25). https://doi.org/10.23913/ride.v13i25.1269

Martínez Bencardino, C. (2019). *Estadística y muestreo* (Décima cuarta edición). Ecoe Ediciones.

Martínez-Garrido, C., & Márquez-Ortiz, J. A. (2024). Desigualdades en el acceso a la educación superior pública. *magis, Revista Internacional de Investigación en Educación*, *17*, 1–22. https://doi.org/10.11144/Javeriana.m17.daes

Mejia González, L., Cujia Berrío, S. E., & Liñan Cuello, Y. I. (2022). Políticas educativas en América Latina: Del modelo economicista a la educación para la sustentabilidad. *Revista Venezolana de Gerencia*, *27*(100), 1489–1501. https://doi.org/10.52080/rvgluz.27.100.13

Miño de Gauto, M. E. (2021). Factores condicionantes de la deserción universitaria. *Ciencia Latina Revista Científica Multidisciplinar*, *5*(4), 5316–5328. https://doi.org/10.37811/cl_rcm.v5i4.691

Montero Caro, M. D. (2021a). Educación, Gobierno Abierto y progreso: Los Objetivos de Desarrollo Sostenible (ODS) en el ámbito educativo. Una visión crítica de la LOMLOE. *Revista de Educación y Derecho*, (23). https://doi.org/10.1344/REYD2021.23.34443

Montero Caro, M. D. (2021b). Educación, Gobierno Abierto y progreso: Los Objetivos de Desarrollo Sostenible (ODS) en el ámbito educativo. Una visión crítica de la LOMLOE. *Revista de Educación y Derecho*, (23). https://doi.org/10.1344/REYD2021.23.34443

Montero Caro, M. D. (2021c). Educación, Gobierno Abierto y progreso: Los Objetivos de Desarrollo Sostenible (ODS) en el ámbito educativo. Una visión crítica de la LOMLOE. *Revista de Educación y Derecho*, (23). https://doi.org/10.1344/REYD2021.23.34443

Moya, E. (2021, enero 1). *Estudio nacional—Summa (2025)*.

Papadogiannis, I., Wallace, M., & Karountzou, G. (2024). Educational Data Mining: A

Foundational Overview. *Encyclopedia*, *4*(4), 1644–1664. https://doi.org/10.3390/encyclopedia4040108

Parra-Sánchez, J. S., Torres Pardo, I. D., & Martínez De Merino, C. Y. (2023). Factores explicativos de la deserción universitaria abordados mediante inteligencia artificial. *Revista Electrónica de Investigación Educativa*, *25*, 1–17. https://doi.org/10.24320/redie.2023.25.e18.4455

Paz-Ruiz, V. (2024). De la ecología a la educación ambiental en la educación primaria en México. *Bio-grafía*, *17*(32), 123–133. https://doi.org/10.17227/bio-grafia.vol.17.num32-20439

Peña, N. M. (2009). Vida universitaria e imaginarios: Posibilidad en definición de políticas sobre educación superior. *Revista Latinoamericana de Ciencias Sociales, Niñez y Juventud*, *7*(1). https://doi.org/10.11600/rlcsnj.7.1.226

Pereira Santana, A. E., & Vidal Cortez, M. (2020). Deserción estudiantil en la educación superior: Reflexiones sobre la gestión enfocada en la retención o la permanencia. *Revista Educación*, 519–533. https://doi.org/10.15517/revedu.v45i1.40602

PNUD. (2022). *ESTADO DE DERECHO FUNDAMENTO DE LA TRANSFORMACIÓN 2022-2030*.

Quintana, Llatasi. (2024). Early Prediction of University Student Dropout Using Machine Learning Models. *Nanotechnology Perceptions*, *20*(S5). https://doi.org/10.62441/nano-ntp.v20iS5.62

Ramírez Díaz, J. A. (2024, julio). *Marcos de políticas para la digitalización de las universidades públicas de México y España | Educación y Ciudad*. https://revistas.idep.edu.co/index.php/educacion-y-ciudad/article/view/3124

Ramírez, N. G., & Castaño, C. A. (2024). Educación inclusiva: Estrategias educativas para el acceso de estudiantes extraedad a la educación superior colombiana. *Revista Boletín Redipe*, *13*(7), 37–52. https://doi.org/10.36260/09wyg678

Ramírez-Díaz, J. A. (2024). Marcos de políticas para la digitalización de las universidades públicas de México y España. *Educación y Ciudad*, (47), e3124–e3124. https://doi.org/10.36737/01230425.n47.2024.3124

Reforma 2019 a los artículos 3°, 31 y 73 de la Constitución Política de los Estados Unidos Mexicanos. (2019). *Perfiles educativos*, *41*(165), 186–208.

https://doi.org/10.22201/iisue.24486167e.2019.165.59496

Resnik, D. B. (2024). *The Ethics of Research with Human Subjects: Protecting People, Advancing Science, Promoting Trust* (Vol. 111). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-82757-0

Rivas, R. A. J. (2023). Transformación Digital de la Universidad de El Salvador: Importancia de una educación hibrida. *Revista Diálogo Interdisciplinario sobre Educación - REDISED*, 197–208.

Rodríguez Gómez, K. (2020). De Progresa-Oportunidades-Prospera a las Becas Benito Juárez: Un análisis preliminar de los cambios en la política social en el sexenio 2018-2024 en México. *Revista Mexicana de Análisis Político y Administración Pública*, *9*(17), 81–91. https://doi.org/10.15174/remap.v9i17.324

Romero, Cristobal, & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, *10*(3), e1355. https://doi.org/10.1002/widm.1355

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, *10*(3), e1355. https://doi.org/10.1002/widm.1355

Ruiz, L. Á. D. (2025). El impacto del comportamiento lector en la educación telesecundaria en la meseta comiteca Tojolabal, Chiapas. *EDUCA. Revista Internacional para la calidad educativa*, *5*(2), 1–37. https://doi.org/10.55040/b1a4jd17

Sánchez, J. C., & Cruz, M. G. G. de la. (2024). Eficacia institucional de los organismos independientes pro-rendición de cuentas: El caso del Instituto Nacional de Transparencia, Acceso a la Información y Protección de Datos Personales (INAI). *Estudios en Derecho a la Información*, 35–59. https://doi.org/10.22201/iij.25940082e.2024.17.18781

Sánchez, R. M. O. (2024). Estudio sobre el Uso de Tecnologías en Educación Superior Antes y Después de Covid- 19. *Ciencia Latina Revista Científica Multidisciplinar*, *8*(2), 2796–2808. https://doi.org/10.37811/cl_rcm.v8i2.10711

Scherer, R., Howard, S. K., Tondeur, J., & Siddiq, F. (2021). Profiling teachers' readiness for online teaching and learning in higher education: Who's ready? *Computers in Human Behavior*, *118*, 106675. https://doi.org/10.1016/j.chb.2020.106675

Sohil, F., Sohali, M. U., & Shabbir, J. (2022). An introduction to statistical learning with

applications in R: By Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, New York, Springer Science and Business Media, 2013, $41.98, eISBN: 978-1-4614-7137-7. *Statistical Theory and Related Fields*, *6*(1), 87–87. https://doi.org/10.1080/24754269.2021.1980261

Sosa, R. M. M. (2022). *ALCANCE ACADÉMICO Y SOCIOECONÓMICO DE BECAS: CASO DE ESTUDIANTES UNIVERSITARIOS UNAH, HONDURA*.

Terraza-Beleño, W. (2019). ESTRATEGIAS DE RETENCIÓN ESTUDIANTIL EN EDUCACIÓNSUPERIOR Y SU INFLUENCIA EN LA DESERCIÓN. *Revista Electrónica en Educación y Pedagogía*, *3*(4).

UNAH. (2023, noviembre 27). *Baja matrícula de estudiantes en todos los niveles tiene como factor común la falta de ingresos económicos*. https://blogs.unah.edu.hn/dircom/baja-matricula-de-estudiantes-en-todos-los-niveles-tiene-como-factor-comun-la-falta-de-ingresos-economicos/

UNAH. (2025a, enero 1). *Consejo de Educación Superior*. https://des.unah.edu.hn/sistema-de-educacion-superior/consejo-de-educacion-superior/?utm_source=chatgpt.com

UNAH. (2025b, abril 7). *Finaliza el Programa de Tutorías para estudiantes con matrícula excepcional—Blogs UNAH*. https://blogs.unah.edu.hn/presencia-universitaria/finaliza-el-programa-de-tutorias-para-estudiantes-con-matricula-excepcional-en-unah?utm_source=chatgpt.com

UNESCO. (s/f). Recuperado el 18 de agosto de 2025, de https://unevoc.unesco.org/home/Dynamic%2BTVET%2BCountry%2BProfiles/country%3DHND?utm_source=chatgpt.com

Universidad de El Salvador. (2022). *Universidad de El Salvador*. https://www.ues.edu.sv/

Urbina-Nájera, A. B., Camino-Hampshire, J. C., & Cruz Barbosa, R. (2020). Deserción escolar universitaria: Patrones para prevenirla aplicando minería de datos educativa. *RELIEVE - Revista Electrónica de Investigación y Evaluación Educativa*, *26*(1). https://doi.org/10.7203/relieve.26.1.16061

Verano, J. P. (2025). La educación superior rural en Colombia: Un estado del arte. *Educación*, *34*(66), 29–48. https://doi.org/10.18800/educacion.202501.A002

Vilchis-Torres, I., & Segura-Lazcano, G. (2025). Adaptación y transformación: Un análisis de la digitalización en las universidades de España y México. *Sociedad & Tecnología*, *8*(S1),

59–71. https://doi.org/10.51247/st.v8iS1.560

Villalobos, C. R. C., Álvarez, R. A. P., & Ramírez, M. E. M. (2023). Diseño instruccional en educación virtual: Migración de cursos de un contexto de aprendizaje presencial a un contexto virtual. *InterSedes*, *24*(50), 312–336. https://doi.org/10.15517/isucr.v24i50.54007

Villar, A., & De Andrade, C. R. V. (2024). Supervised machine learning algorithms for predicting student dropout and academic success: A comparative study. *Discover Artificial Intelligence*, *4*(1), 2. https://doi.org/10.1007/s44163-023-00079-z

Villegas-Ch, W., Govea, J., & Revelo-Tapia, S. (2023). Improving Student Retention in Institutions of Higher Education through Machine Learning: A Sustainable Approach. *Sustainability*, *15*(19), 14512. https://doi.org/10.3390/su151914512

Yagual, C. A. R., Tigua, D. D. P., & Mullo, E. C. (2022). Brecha digital y educación virtual en estudiantes de secundaria de una institución educativa ecuatoriana. *Ciencia Latina Revista Científica Multidisciplinar*, *6*(6), 12738–12749. https://doi.org/10.37811/cl_rcm.v6i6.4279

Zárate-Valderrama, J., Bedregal-Alpaca, N., & Cornejo-Aparicio, V. (2021). Modelos de clasificación para reconocer patrones de deserción en estudiantes universitarios. *Ingeniare. Revista chilena de ingeniería*, *29*(1), 168–177. https://doi.org/10.4067/S0718-33052021000100168

Zavala, M. D. L. B., Conde, M. G. J., Hernández, J. D. M., & Valencia, C. P. (2025). Dimensiones de valor del patrimonio cultural inmaterial en los municipios de la región capital de Veracruz, México 2020-2024. *UVserva*, (19), 98–111. https://doi.org/10.25009/uvs.vi19.3123