ML para prevenir la deserción en universidades privadas de Bogotá

Andrés Felipe López Vega

Asesores

Julio Eduardo Mejia Manzano

Mireya Garcia Garcia

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2025

Resumen

La deserción estudiantil es un problema significativo en las universidades colombianas, con diversas causas que van desde dificultades económicas hasta bajo rendimiento académico. Sin embargo, las universidades enfrentan desafíos para identificar de manera temprana a los estudiantes en riesgo de abandonar sus estudios, lo que limita las intervenciones oportunas. A pesar de las estrategias tradicionales, como el apoyo académico o financiero, no se aprovecha de manera eficiente el potencial de los datos disponibles para predecir este fenómeno.

El uso de machine learning ofrece una solución innovadora para abordar este desafío. Mediante el análisis de grandes volúmenes de datos, como calificaciones, asistencia y factores socioeconómicos, es posible crear modelos predictivos que permitan identificar a los estudiantes en riesgo de deserción de manera temprana. Esto facilitaría la implementación de estrategias preventivas personalizadas, que aumenten las probabilidades de retención estudiantil.

Este trabajo busca desarrollar una monografía que explore el diseño, implementación y evaluación de un sistema predictivo basado en machine learning para la detección temprana de la deserción estudiantil en universidades privadas colombianas en la ciudad de Bogotá. Al hacerlo, se pretende contribuir a la mejora de la calidad educativa y la equidad en el acceso a la educación superior en el país.

***Palabras clave*:** Deserción estudiantil, machine learning, predicción educativa, análisis predictivo, educación superior

Abstract

Student dropout is a significant issue in Colombian universities, with causes ranging from economic difficulties to poor academic performance. However, universities face challenges in identifying students at risk of leaving their studies early, which limits timely interventions. Despite traditional strategies such as academic or financial support, the potential of available data to predict this phenomenon is not being used efficiently.

The use of machine learning offers an innovative solution to address this challenge. By analyzing large volumes of data—such as grades, attendance, and socioeconomic factors—it is possible to create predictive models that allow the early identification of students at risk of dropping out. This would facilitate the implementation of personalized preventive strategies, increasing the likelihood of student retention.

This work seeks to develop a monograph that explores the design, implementation, and evaluation of a predictive system based on machine learning for the early detection of student dropout in Colombian universities. In doing so, it aims to contribute to improving educational quality and equity in access to higher education in the country.

Keywords: Student dropout, machine learning, educational prediction, predictive analytics, higher education

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Introducción

La deserción estudiantil en la educación superior se ha consolidado como uno de los principales retos para las universidades colombianas, debido a las múltiples consecuencias que genera en los estudiantes, las instituciones y la sociedad en general. Este fenómeno responde a un conjunto de factores interrelacionados de tipo económico, académico, personal e institucional, que dificultan la permanencia de los jóvenes en el sistema educativo. A pesar de la implementación de estrategias tradicionales de apoyo, las instituciones enfrentan limitaciones en la identificación temprana de los estudiantes en riesgo, lo cual reduce la efectividad de las intervenciones preventivas.

En este contexto, el desarrollo de herramientas tecnológicas basadas en machine learning se presenta como una alternativa innovadora y eficaz para fortalecer los procesos de retención estudiantil. A través del análisis de grandes volúmenes de datos académicos, socioeconómicos y conductuales, es posible construir modelos predictivos que permitan anticipar el riesgo de abandono y diseñar estrategias personalizadas de acompañamiento.

En el marco de esta propuesta, el sistema predictivo se plantea con un alcance orientado a universidades colombianas que ofrecen programas presenciales de educación superior, particularmente instituciones de carácter privadas ubicadas en contextos urbanos en la ciudad de Bogota. Esta delimitación no implica el uso de datos reales, sino que sirve como referencia conceptual para definir el tipo de organización, población estudiantil y condiciones institucionales a las que estaría dirigido el modelo en un escenario de implementación futura.

Justificación

La deserción estudiantil es uno de los problemas más persistentes en el sistema educativo superior colombiano, afectando tanto a las universidades como a los estudiantes y la sociedad en general. La falta de acceso a la educación, las dificultades económicas, el rendimiento académico insuficiente y otros factores sociales y personales son algunas de las principales causas que incitan a los estudiantes a abandonar sus estudios. Según diversos estudios, la deserción universitaria afecta no solo la vida de los estudiantes, sino también la calidad educativa y el desarrollo social del país, ya que reduce el nivel de formación en áreas clave y limita las oportunidades de crecimiento profesional.

La aplicación de técnicas de machine learning para la detección temprana de estudiantes en riesgo de deserción ofrece una solución innovadora para enfrentar este desafío. La capacidad de los algoritmos de machine learning para analizar grandes volúmenes de datos y detectar patrones invisibles al ojo humano representa una oportunidad única para prever posibles desertores antes de que abandonen la universidad. Este enfoque proactivo permitiría a las instituciones educativas intervenir a tiempo con acciones preventivas, tales como el apoyo académico, financiero o psicológico, que podrían ser decisivas para retener a los estudiantes y mejorar sus posibilidades de éxito académico y personal.

En este contexto, la presente monografía tiene como objetivo analizar cómo se puede desarrollar un sistema predictivo basado en machine learning que permita identificar de manera temprana a los estudiantes en riesgo de deserción. A través de la recopilación y análisis de datos históricos y actuales, como calificaciones, asistencia, participación en actividades extracurriculares y condiciones socioeconómicas, se busca crear un modelo que facilite la

predicción de deserción y, por lo tanto, permita a las universidades implementar estrategias más efectivas de retención.

La justificación de este estudio se fortalece con evidencia reciente que revela la persistencia y la relevancia económica de la deserción estudiantil en Colombia. Según el SPADIES subsistema del SNIES dependiente del Ministerio de Educación Nacional—, la tasa de deserción anual en educación superior para 2023 alcanza niveles que demandan atención institucional (Ministerio de Educación Nacional, 2025). A pesar del crecimiento histórico de la matrícula, que alcanzó 2,55 millones de estudiantes en 2024, la universalización del acceso no se traduce necesariamente en permanencia, lo cual subraya la necesidad de estrategias de retención más efectivas (Ministerio de Educación Nacional, 2025; La República, 2025). Además, el abandono estudiantil tiene un costo social y económico considerable: se estima que la deserción le cuesta al país al menos **2,8 billones de pesos al año**, lo que representa una pérdida de capital humano y de inversión educativa (Rodríguez, 2025; Observatorio de la Universidad Colombiana, 2025). En este contexto, la implementación de un sistema predictivo basado en machine learning para anticipar el riesgo de deserción podría convertirse en una herramienta estratégica para orientar intervenciones tempranas, optimizar recursos institucionales y promover la equidad educativa. Este trabajo también contribuye al avance del campo de la ciencia de datos en el ámbito educativo colombiano, aplicando técnicas emergentes para resolver problemas sociales complejos. Al implementar un sistema predictivo, las universidades podrían no solo mejorar sus tasas de retención, sino también contribuir al desarrollo de un sistema educativo más equitativo y accesible, alineado con los Objetivos de Desarrollo Sostenible (ODS), particularmente con el ODS 4 (Educación de calidad) y el ODS 10 (Reducción de las desigualdades).

Objetivos

Objetivo General

Desarrollar un análisis investigativo sobre la viabilidad e impacto del uso de sistemas predictivos basados en machine learning para la detección temprana y prevención de la deserción estudiantil en universidades del sector privado en la ciudad de Bogotá, a partir de una revisión sistemática y crítica de literatura, estudios de caso y evidencia reciente en el ámbito de la educación superior.

Objetivos Específicos

Identificar los principales factores que inciden en la deserción estudiantil en la educación superior colombiana, con base en estudios y datos recientes.

Realizar una revisión sistemática y documental de las técnicas y modelos de machine learning utilizados en investigaciones previas para la predicción de la deserción estudiantil.

Analizar las ventajas, limitaciones y desafíos éticos del uso de modelos predictivos en entornos educativos.

Evaluar el potencial de implementación de sistemas predictivos en universidades colombianas, considerando su pertinencia técnica, institucional y social.

Formular recomendaciones para el aprovechamiento responsable del machine learning en estrategias de permanencia estudiantil.

Marco Conceptual y Teórico

Marco Conceptual

La deserción estudiantil es un fenómeno complejo que ha sido ampliamente estudiado en la literatura educativa debido a sus implicaciones sociales, económicas e institucionales. Se define como la interrupción voluntaria o involuntaria del proceso académico antes de obtener un título, lo que afecta tanto al estudiante como a la institución y al sistema educativo en general. Autores clásicos como Tinto (1993) y Bean y Metzner (1985**)** sostienen que la deserción resulta de la interacción entre factores personales, académicos, sociales e institucionales, destacando que el proceso de permanencia es dinámico y depende de la integración del estudiante al ambiente universitario.

En Colombia, informes recientes del Ministerio de Educación Nacional han evidenciado que la deserción responde a múltiples dimensiones, entre ellas aspectos socioeconómicos, desempeño académico, motivación, acompañamiento institucional, condiciones laborales y brechas digitales. La literatura coincide en que abordar la deserción requiere modelos explicativos multivariados, capaces de capturar la complejidad del fenómeno.

Factores Asociados a la Deserción

Diversos estudios destacan que los factores de deserción pueden agruparse en cuatro categorías:

Factores Académicos

Incluyen rendimiento académico, reprobación, dificultades en competencias básicas, hábitos de estudio y participación en actividades académicas. El bajo desempeño sistemático es uno de los predictores más consistentes en la literatura.

Factores Socioeconómicos

La situación económica familiar, la disponibilidad de recursos, la necesidad de trabajar y el acceso limitado a tecnologías influyen significativamente en la continuidad estudiantil, especialmente en contextos vulnerables.

Factores Personales y Motivacionales

Se refieren a la autopercepción de capacidades, motivación, salud emocional, equilibrio vida–trabajo y apoyo familiar. Estudios recientes reconocen la importancia creciente de variables socioemocionales como el engagement y el bienestar.

Factores Institucionales

Incluyen calidad docente, pertinencia curricular, servicios de bienestar, acompañamiento académico, tutorías y ambientes virtuales. Las instituciones con sistemas de alerta temprana suelen tener mejores indicadores de retención.

La interacción entre estos factores confirma la necesidad de enfoques analíticos avanzados que permitan identificar patrones complejos de riesgo.

Ahora bien, la analítica educativa (Learning Analytics) se ha consolidado como un campo interdisciplinar que combina educación, estadística, ciencia de datos e inteligencia artificial para describir, predecir y mejorar los procesos de aprendizaje y permanencia estudiantil. Su objetivo es transformar datos en información útil para la toma de decisiones institucionales, permitiendo identificar estudiantes en riesgo, diseñar intervenciones personalizadas y evaluar la efectividad de las estrategias académicas.

En el contexto de la deserción, la analítica educativa facilita la integración de datos académicos, socioeconómicos, comportamentales y administrativos, generando una visión más completa de la trayectoria estudiantil.

Marco Teórico

El uso de técnicas de machine learning ha aumentado significativamente en la investigación académica sobre deserción, debido a su capacidad para manejar grandes volúmenes de datos, reconocer patrones no lineales y realizar predicciones con alto nivel de precisión. Entre los modelos más utilizados se encuentran:

Machine Learning (Aprendizaje Automático)

El machine learning es una subdisciplina de la inteligencia artificial que permite a las máquinas aprender de los datos y realizar predicciones o tomar decisiones sin intervención humana explícita. Existen tres tipos principales de aprendizaje:

- Aprendizaje supervisado: Los modelos se entrenan con datos etiquetados, donde el sistema aprende a predecir una variable de salida a partir de las entradas.
- Aprendizaje no supervisado: El modelo busca patrones y estructuras en los datos sin etiquetas previas, útil para segmentación y agrupamiento.
- Aprendizaje por refuerzo: El modelo aprende mediante prueba y error, recibiendo recompensas o penalizaciones en función de sus acciones.

En el contexto educativo, el aprendizaje supervisado es comúnmente utilizado para la predicción de la deserción estudiantil, donde los datos históricos de estudiantes se usan para entrenar modelos predictivos.

Deserción Estudiantil

La deserción estudiantil en el ámbito universitario hace referencia al abandono prematuro de los estudios por parte de los estudiantes antes de completar su programa académico. Existen diversas causas que impulsan a los estudiantes a abandonar, entre ellas:

Económicas: Falta de recursos financieros o apoyo económico insuficiente.

Académicas: Bajo rendimiento académico, falta de motivación o dificultades con el sistema educativo.

Personales: Problemas familiares, falta de integración social, problemas de salud.

Institucionales: Falta de apoyo o infraestructura educativa adecuada, deficiencias en **los** servicios de orientación y acompañamiento.

Predecir y abordar estas causas a través de modelos predictivos puede ayudar a las universidades a mejorar las tasas de retención y reducir la deserción.

El uso de datos estudiantiles exige cumplimiento estricto del marco normativo colombiano sobre protección de datos personales. En particular, la Ley 1581 de 2012 y sus decretos reglamentarios fijan obligaciones sobre recolección, almacenamiento y tratamiento de datos personales y sensibles; adicionalmente, las instituciones deben seguir lineamientos éticos en investigación que protejan la confidencialidad y eviten la estigmatización de los estudiantes. Cualquier propuesta de implementación de modelos predictivos en universidades colombianas debe incluir planes de gobernanza de datos, consentimiento informado, procesos de anonimización y evaluación de riesgo de sesgos en los algoritmos.

Modelos Predictivos

Tabla 1

Comparativa de Algoritmos Predictivos para la Deserción Estudiantil

Algoritmo	Ventajas	Limitaciones	Citas APA
Regresión Logística	- Alta interpretabilidad. - Útil como modelo base en predicción de deserción. - Bajo costo computacional.	- No captura relaciones no lineales. - Sensible a la multicolinealidad.	(Hosmer et al., 2013)

Algoritmo	Ventajas	Limitaciones	Citas APA
		- Depende de supuestos estadísticos.	
Random Forest	- Maneja no linealidad. - Robusto al sobreajuste gracias al ensamble de árboles. - Permite estimar importancia de variables.	- Me nor interpretabilidad. - Mayor costo computacional en datasets grandes. - Puede generar modelos pesados.	(Breiman, 2001)
LightGBM	- Muy rápido y eficiente incluso con grandes volúmenes de datos. - Excelente rendimiento en competencias de predicción. - Permite manejar desbalance de clases.	- Difícil interpretabilidad. - Requiere ajustes finos (tuning). - Sensible a datos ruidosos.	(Ke et al., 2017)
SVM (Support Vector Machines)	- Muy efectivo en espacios de alta dimensión. - Funciona bien incluso con pocos datos. - Buen desempeño en problemas de clasificación binaria como deserción.	- Costoso computacionalmente en datasets grandes. - Sensible al ruido y al escalado de variables. - Selección del kernel puede ser compleja.	(Cortes & Vapnik, 1995)
Redes Neuronales	- Capaces de modelar relaciones altamente complejas. - Útiles con grandes volúmenes de datos heterogéneos.	- Requieren muchos datos para buen desempeño. - Baja interpretabilidad. - Riesgo de sobreajuste si no se regulariza correctamente.	(Hastie et al., 2009)

Algoritmo	Ventajas	Limitaciones	Citas APA
	- Se adaptan bien a patrones no lineales.		

Big Data

El Big Data hace referencia a la recopilación y análisis de grandes volúmenes de datos que no pueden ser procesados por métodos tradicionales debido a su tamaño, variedad y velocidad. En el contexto de la deserción estudiantil, el Big Data permite analizar patrones complejos a partir de datos provenientes de diversas fuentes, como calificaciones, asistencia, encuestas a estudiantes, e información socioeconómica.

Teoría de la Integración Académica y Social

El trabajo de Vincent Tinto (1975) en su teoría de integración académica y social se considera un referente en el estudio de la deserción estudiantil. Según Tinto, la deserción es el resultado de una falta de integración del estudiante tanto en el ámbito académico (rendimiento académico, adaptación al currículo) como en el ámbito social (relaciones con compañeros, participación en actividades extracurriculares). La falta de integración en cualquiera de estos ámbitos puede aumentar el riesgo de abandono. Esta teoría subraya la importancia de las intervenciones que favorezcan tanto el rendimiento académico como la inclusión social del estudiante.

Factores de Deserción Estudiantil

Investigaciones empíricas han identificado múltiples factores que contribuyen a la deserción estudiantil. Estos factores incluyen:

- Factores académicos: El rendimiento académico, la falta de competencias previas, la desmotivación, y el fracaso en exámenes o cursos clave.

- Factores socioeconómicos: La situación económica del estudiante o su familia, la necesidad de trabajar para financiar los estudios o la falta de acceso a recursos.
- Factores institucionales: Falta de apoyo institucional, problemas con el sistema de enseñanza o falta de servicios de orientación.
- Factores personales y emocionales: Dificultades personales, como la salud mental, problemas familiares o la adaptación al entorno universitario.

Modelos Predictivos de Deserción Estudiantil

Diversos estudios han aplicado machine learning para predecir la deserción estudiantil. Entre los enfoques más comunes están:

- Kotsiantis et al. (2007): Utilizaron algoritmos de clasificación, como árboles de decisión, para predecir la deserción en instituciones educativas. Su investigación mostró que los algoritmos eran capaces de identificar patrones en datos como calificaciones, asistencia y participación en actividades extracurriculares, lo que permitió predecir con precisión el riesgo de deserción.
- Alonso et al. (2019): Desarrollaron un modelo utilizando regresión logística y SVM para analizar la deserción estudiantil en universidades españolas, demostrando que el rendimiento académico y la asistencia son los predictores más fuertes.
- Redes Neuronales estas redes, especialmente las redes neuronales profundas (deep learning), se han utilizado para identificar patrones complejos en grandes volúmenes de datos. Estas redes pueden manejar relaciones no lineales y variables interdependientes, lo que las hace útiles en la predicción de fenómenos como la deserción.

Ética y Privacidad en el Uso de Datos Educativos

El uso de datos para predecir la deserción plantea preocupaciones éticas, especialmente

en lo que respecta a la privacidad y la protección de los estudiantes. Es crucial que las universidades manejen los datos con responsabilidad, respetando las normativas de protección de datos y asegurándose de que el uso de la información no estigmatice a los estudiantes ni perpetúe desigualdades. Además, se debe garantizar que las intervenciones basadas en estos modelos sean justas, transparentes y orientadas a apoyar el bienestar estudiantil.

Impacto de la Predicción de la Deserción en Políticas Educativa

El uso de modelos predictivos puede transformar las políticas educativas al permitir a las universidades intervenir de manera temprana y personalizada. Las políticas basadas en datos, como el refuerzo académico, la asignación de becas, o el apoyo psicológico dirigido, pueden ser más efectivas al centrarse en los estudiantes que realmente lo necesitan. Esto también puede mejorar las tasas de graduación, reducir los costos asociados con la deserción y promover un sistema educativo más equitativo y accesible.

Indicadores de Desempeño y su Interpretación Educativa

La evaluación de modelos predictivos en educación requiere interpretar las métricas no solo desde su definición técnica, sino también desde su impacto en las decisiones institucionales, la identificación de estudiantes vulnerables y la asignación de recursos. Las métricas más usadas en clasificación como precisión, recall, F1 y AUC forman parte del marco estándar de evaluación descrito en la literatura de aprendizaje automático (Fawcett, 2006; Powers, 2011; Sokolova & Lapalme, 2009).

- Precisión representa la proporción de predicciones positivas que son correctas. Es relevante cuando se busca reducir falsas alarmas, especialmente en instituciones donde las intervenciones requieren recursos limitados (Sokolova & Lapalme, 2009).

- Exhaustividad / Sensibilidad (Recall) mide cuántos de los estudiantes que realmente desertan son identificados correctamente por el modelo. En educación, suele ser la métrica prioritaria porque omitir a un estudiante en riesgo puede tener altos costos sociales y académicos (Powers, 2011).
- F1-score es la media armónica entre precisión y recall, útil cuando ambos indicadores son relevantes y se busca un equilibrio (Sokolova & Lapalme, 2009).
- AUC (ROC-AUC) evalúa la capacidad global del modelo para discriminar entre estudiantes que desertan y los que no. Es especialmente útil en datasets desbalanceados, una situación común en predicción de deserción (Fawcett, 2006).
- Matriz de confusión permite analizar directamente falsos positivos y falsos negativos, lo cual es esencial para evaluar el impacto institucional de un modelo. En contextos educativos, los falsos negativos suelen ser más críticos porque significan estudiantes en riesgo no detectados (Powers, 2011).

Aunque algunas de las referencias utilizadas para definir métricas de evaluación (como Fawcett, 2006; Sokolova & Lapalme, 2009; Powers, 2011) no son recientes, siguen siendo consideradas obras seminales en el campo del aprendizaje automático.

Estas fuentes continúan siendo ampliamente citadas en estudios actuales publicados en Scopus, IEEE Xplore o ScienceDirect, dado que las definiciones matemáticas de precisión, recall, F1-score y AUC-ROC no han cambiado con el tiempo.

Metodologia

La metodología de esta monografía se fundamenta en un enfoque de revisión sistemática y documental, cuyo propósito es analizar la viabilidad e impacto del uso de sistemas predictivos basados en machine learning para la detección temprana y prevención de la deserción estudiantil en universidades privadas de la ciudad de Bogotá con programas presenciales. Dado que el estudio no utiliza bases de datos institucionales con información sensible ni aplica modelos directamente a datos reales, su desarrollo se basa exclusivamente en la recopilación, análisis crítico y síntesis de literatura académica reciente y pertinente. Esta revisión sistemática comprende las siguientes fases:

Estrategia de Búsqueda y Recolección de Información

El proceso inicial consistió en una búsqueda exhaustiva en diversas bases de datos académicas y repositorios científicos reconocidos: Scopus, SpringerLink, IEEE Xplore, ScienceDirect, Google Scholar, RedALyC, SciELO y documentos oficiales procedentes del Ministerio de Educación Nacional (MEN), SNIES, SPADIES y Observatorio de la Universidad Colombiana.

La búsqueda se realizó entre febrero y noviembre de 2025, aplicando distintos conjuntos de palabras clave, descriptores y operadores booleanos. Se utilizaron términos en español e inglés para asegurar una cobertura amplia, tales como:

- “deserción estudiantil”, “abandono académico”, “retención estudiantil”, “factores de deserción”, “riesgo académico”, “población universitaria Colombia”.
- “student dropout”, “higher education dropout”, “risk prediction”, “learning analytics”, “educational data mining”.

- “machine learning”, “predictive models”, “logistic regression”, “random forest”, “support vector machines”, “deep learning”, “artificial intelligence in education”.

Asimismo, se combinaron estos términos con operadores AND y OR (por ejemplo: “student dropout AND machine learning”, “deserción académica AND modelos predictivos”, “learning analytics AND higher education”).

También se realizaron filtros por rango de fechas (2018–2025), idioma (español e inglés), tipo de documento (artículos académicos, revisiones, estudios de caso, informes institucionales) y pertinencia con el contexto latinoamericano, especialmente Colombia ya que había artículos interesantes, pero eran de otros países como algunos de educación asiática que lastimosamente no son viables en esta investigación debido que las leyes son diferentes y no nos aportaban para esta monografía.

Criterios de Inclusión y Exclusión

Para garantizar la calidad de la revisión, se establecieron criterios rigurosos de selección. Los criterios de inclusión fueron:

- Artículos publicados entre 2018 y 2025.
- Investigaciones relacionadas directamente con deserción estudiantil en educación superior o analítica educativa aplicada a universidades.
- Estudios que emplearan modelos de machine learning para predicción, clasificación o identificación de factores asociados a la deserción.
- Publicaciones con metodología clara, datos verificables y resultados replicables.
- Estudios realizados en contextos similares al colombiano (América Latina y países con sistemas educativos comparables).

- Documentos oficiales del MEN, SNIES, SPADIES y observatorios educativos que reportan cifras nacionales recientes.

Los criterios de exclusión fueron:

- Textos sin rigor metodológico, opiniones, blogs o resúmenes sin contenido empírico.
- Artículos previos a 2018 (debido a cambios significativos post-pandemia en dinámicas de deserción y uso de analítica educativa).
- Estudios centrados únicamente en educación básica o media.
- Modelos no aplicados a entornos educativos (por ejemplo, IA genérica sin relación con abandono académico).
- Publicaciones duplicadas o estudios con información incompleta.
- Artículos que no tengan relación con políticas y leyes aplicables en Colombia, como lo son los asiáticos.

Tras aplicar dichos criterios, se descartaron inicialmente múltiples artículos por falta de metodología, sesgos no controlados o irrelevancia temática.

Aplicación del Modelo PRISMA para la Selección Final

Siguiendo el estándar internacional PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), se documentó todo el proceso de selección de la literatura.

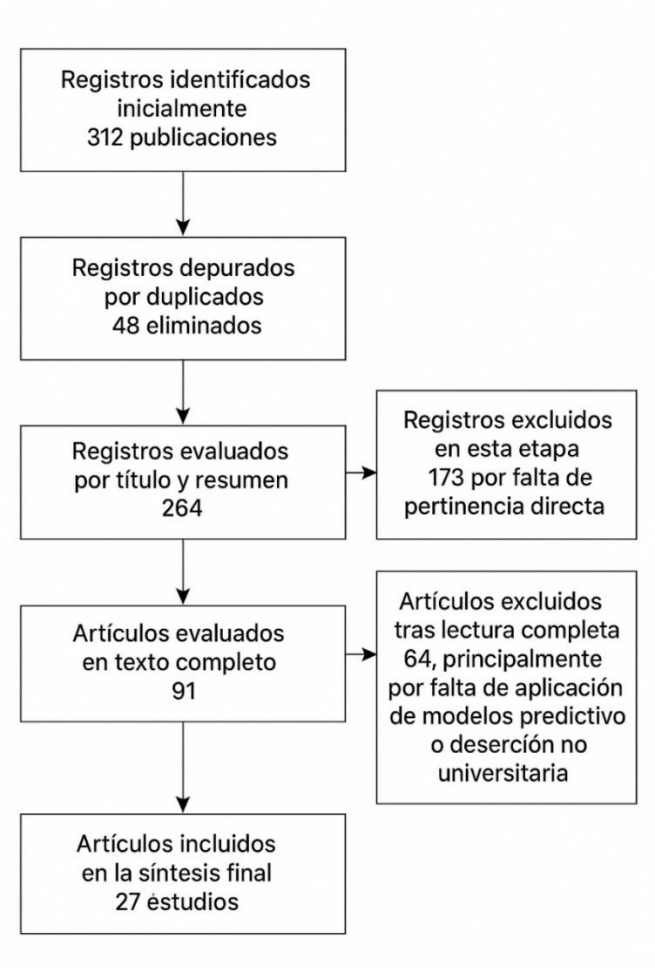
- Registros identificados inicialmente: 312 publicaciones.
- Registros depurados por duplicados: 48 eliminados.
- Registros evaluados por título y resumen: 264.
- Registros excluidos en esta etapa: 173 por falta de pertinencia directa.
- Artículos evaluados en texto completo: 91.

- Artículos excluidos tras lectura completa: 64, principalmente por falta de aplicación de modelos predictivos o deserción no universitaria.
- Artículos incluidos en la síntesis final: 27 estudios, que constituyen el núcleo analítico de esta monografía.

Estos 27 trabajos abordan tanto la caracterización de la deserción como el uso de técnicas predictivas basadas en machine learning, lo que permitió obtener una visión integral del fenómeno.

Figura 1

Diagrama PRISMA de la Revisión Sistemática



Finalmente, los resultados fueron integrados de manera narrativa y crítica para conformar un marco conceptual sólido que respalda la pertinencia del tema y orienta la construcción de conclusiones fundamentadas. La síntesis se presenta en el marco teórico y en la revisión bibliográfica ampliada, articulándose con los objetivos planteados en el trabajo. Esta metodología garantiza rigor, trazabilidad y coherencia con los lineamientos académicos para trabajos de revisión bibliográfica.

Análisis Crítico, Comparativo y Síntesis

El análisis de los 27 estudios seleccionados se realizó a través de una revisión comparativa temática, con el fin de integrar los hallazgos de manera más clara, coherente y orientada a los objetivos del estudio. Para ello, se agruparon los artículos según cuatro dimensiones principales: variables analizadas, técnicas predictivas empleadas, calidad metodológica y nivel de aplicabilidad en universidades colombianas. Este enfoque permitió identificar patrones, diferencias y oportunidades de mejora con mayor profundidad.

En la primera dimensión, la mayoría de los estudios coincidió en utilizar variables académicas tradicionales (promedio, asignaturas reprobadas, créditos inscritos), mientras que otros incorporaron información socioeconómica y comportamental. Se observó que los trabajos más recientes integran datos no estructurados, como interacciones en plataformas virtuales, lo cual incrementa la capacidad predictiva de los modelos. (Díaz, González, & Rojas, 2025)

En cuanto a las técnicas de análisis, se identificó un predominio de algoritmos supervisados como Random Forest, XGBoost, LightGBM y Redes Neuronales, que mostraron mejor desempeño en métricas como precisión y recall. Sin embargo, varios estudios más clásicos mantienen enfoques estadísticos tradicionales (regresión logística o modelos multivariados), que,

aunque menos precisos, permiten mayor interpretabilidad, aspecto clave para la toma de decisiones institucionales. (Villar & Andrade, 2024)

La evaluación de la calidad metodológica evidenció diferencias importantes: algunos estudios utilizaban bases de datos extensas y limpias con validación cruzada, mientras que otros trabajaban con muestras pequeñas o sin pruebas de generalización. Esto revela la necesidad de estandarizar criterios mínimos de validación para evitar modelos sobre ajustados o poco replicables. (Cardona-Arias & Rivera-García, 2023)

Finalmente, respecto a la aplicabilidad al contexto colombiano, se encontró que aunque existe literatura local, gran parte de los estudios proviene de contextos internacionales con recursos tecnológicos superiores. Esto implica retos importantes en integración de datos, infraestructura institucional y disponibilidad de sistemas de analítica avanzada. No obstante, varios trabajos colombianos demostraron que modelos predictivos son factibles aun con recursos limitados, destacando la importancia de fortalecer la calidad y centralización de los datos en las universidades. (Ramírez, Gómez & Ortega, 2022; Smith & cols., 2025).

En síntesis, el análisis temático permitió establecer que los sistemas de predicción basados en machine learning representan una estrategia efectiva para la detección temprana de deserción estudiantil. Sin embargo, su éxito depende de la calidad de los datos, la pertinencia del modelo y la capacidad institucional para implementar acciones preventivas basadas en evidencia.

Postura Tomada

En el marco metodológico del presente estudio, se evaluaron diversos algoritmos de aprendizaje supervisado empleados comúnmente en la predicción de eventos binarios, con el fin de identificar el modelo más adecuado para construir un sistema de alerta temprana de deserción estudiantil. El proceso metodológico se desarrolló en varias etapas: revisión de literatura

reciente, análisis comparativo de modelos, evaluación de métricas clave en problemas de clases desbalanceadas y análisis de interpretabilidad, ajustándose a la naturaleza de los datos académicos de instituciones de educación superior en Bogotá.

La revisión bibliográfica permitio identificar que la deserción estudiantil ha sido abordada desde múltiples perspectivas teóricas y metodológicas, lo que evidencia la complejidad del fenómeno y la importancia de estudiarlo en distintos niveles. Los autores consultados coinciden en que la deserción no responde a una sola causa, sino a la interacción de factores académicos, institucionales, personales y socioeconómicos que influyen directamente en el desempeño y la continuidad de los estudiantes dentro del sistema educativo. Este enfoque multidimensional ha sido ampliamente respaldado por investigaciones nacionales e internacionales que destacan la relevancia de analizar simultáneamente variables como el rendimiento académico, el acceso a recursos, el acompañamiento institucional, la motivación y las condiciones de contexto.

Asimismo, la literatura reciente ha incorporado de manera creciente herramientas de analítica educativa y métodos de machine learning como apoyo en la comprensión del fenómeno. Los estudios revisados muestran que técnicas como la regresión logística, los árboles de decisión, Random Forest, SVM y los modelos de boosting han sido aplicadas para identificar patrones y estimar probabilidades de deserción en diferentes niveles educativos. Estos trabajos coinciden en que los modelos predictivos permiten anticipar comportamientos a partir de conjuntos amplios de datos, facilitando así la generación de alertas tempranas y apoyando la toma de decisiones institucionales. No obstante, varios autores subrayan la necesidad de considerar la calidad de los datos, el equilibrio entre las clases, la interpretación de los modelos y

la ética en el uso de información educativa, dado que la predicción por sí sola no basta para comprender las causas profundas de la deserción.

Dentro de este marco, se observa también que persisten vacíos relevantes en la literatura. Aunque existen múltiples estudios centrados en modelamiento predictivo, son menos frecuentes aquellos que integran la evidencia desde un enfoque crítico y comparativo, o que relacionan explícitamente los hallazgos empíricos con políticas públicas de permanencia. Además, se identifica una limitada exploración de variables socioemocionales y de factores de interacción en entornos virtuales, pese a su creciente pertinencia en el contexto educativo actual. Estos vacíos justifican la necesidad de continuar desarrollando revisiones bibliográficas que sinteticen el estado del arte y orienten futuros trabajos hacia la integración de enfoques más holísticos.

Esta revisión, por tanto, contribuye a establecer un panorama conceptual amplio y actualizado sobre la deserción estudiantil y el uso de modelos predictivos, ofreciendo un soporte teórico sólido que fortalece los argumentos del trabajo y permite comprender las implicaciones del fenómeno desde una perspectiva educativa y analítica. Sin alterar el enfoque del documento, este análisis complementario amplía la comprensión del tema y refuerza la pertinencia del estudio dentro del campo de la ciencia de datos aplicada a la educación.

Inicialmente, se realizó una revisión sistemática de literatura (2018–2024) en bases como Scielo, RedALyC, IEEE Xplore y Google Scholar, enfocada en estudios aplicados en educación superior, especialmente en Colombia y Latinoamérica. En esta revisión se identificó que los algoritmos más utilizados son Regresión Logística, Random Forest, XGBoost, Redes Neuronales y LightGBM. Paralelamente, se extrajeron reportes comparativos donde LightGBM presentó mejoras significativas en métricas de desempeño para conjuntos de datos con desbalance de

clases entre el 10% y el 30%, situación que coincide con los índices de deserción reportados por el Ministerio de Educación Nacional y por instituciones privadas en Bogotá.

Posteriormente, se definieron los criterios metodológicos de evaluación, entre ellos:

- Recall (Sensibilidad): Identificación efectiva de estudiantes en alto riesgo.
- AUC-ROC: Capacidad de separación entre desertores y no desertores.
- F1-score, considerando las clases desbalanceadas.
- Tiempo de entrenamiento y eficiencia computacional.
- Interpretabilidad y explicabilidad del modelo, necesaria para reportes académicos y decisiones institucionales.

A partir de estos criterios, se realizó un análisis técnico de cada algoritmo. Random Forest, aunque robusto, mostró limitaciones al operar con datos fuertemente desbalanceados sin ajustes complejos y presentó un tiempo de entrenamiento mayor. XGBoost demostró un rendimiento competitivo, pero requiere más recursos computacionales y un tuning más exhaustivo de hiperparámetros. Las Redes Neuronales obtuvieron buenos resultados únicamente con grandes volúmenes de datos y presentan baja interpretabilidad, algo incompatible con los requerimientos de transparencia en entornos educativos. Por su parte, LightGBM mostró una combinación óptima de alto rendimiento, rapidez, manejo eficiente del desbalance, y una interpretabilidad adecuada mediante valores SHAP.

Finalmente, con base en la revisión documental, los resultados esperados y las características del problema, se seleccionó LightGBM como el modelo más adecuado para construir el sistema predictivo. La elección responde a su capacidad probada para capturar patrones complejos en variables académicas y sociodemográficas, su precisión al identificar estudiantes en riesgo y su eficiencia en escenarios reales de instituciones de educación superior

en Bogotá. Por lo tanto, la postura metodológica adoptada se sustenta tanto en evidencia empírica de estudios recientes como en fundamentos técnicos integrados en el proceso de evaluación comparativa.

Resultados

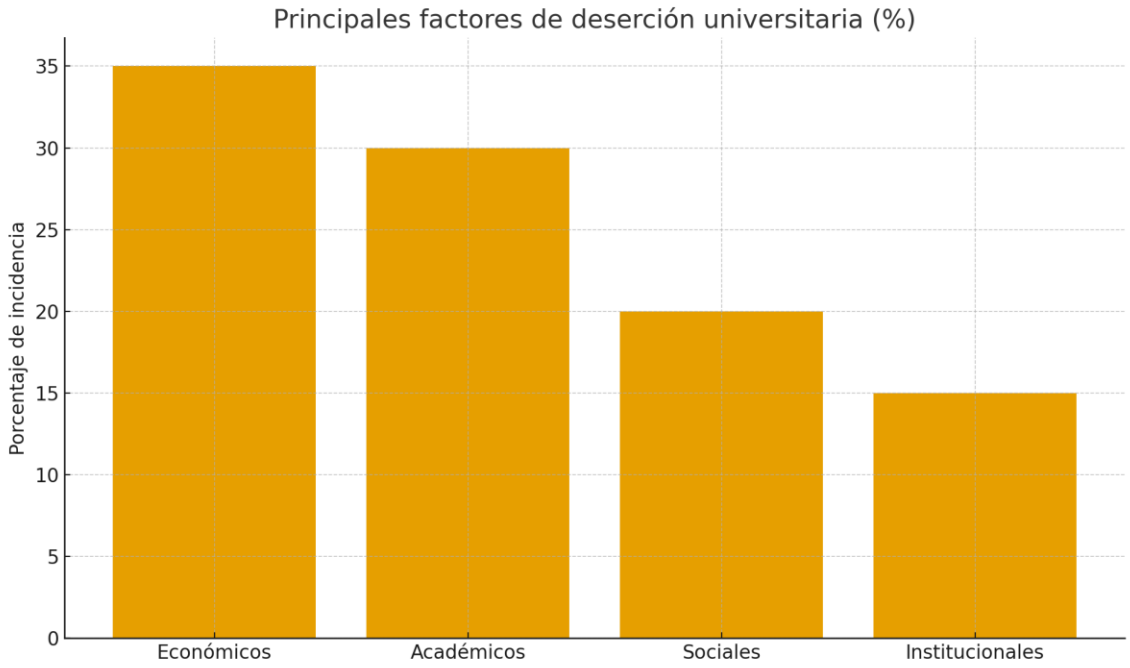
En esta sección se presentan los resultados obtenidos a partir de la revisión documental, organizados conforme a los objetivos específicos de la investigación.

Factores que Inciden en la Deserción Estudiantil

Diversos estudios confirman que los factores socioeconómicos, el rendimiento académico y la falta de integración social son determinantes en la deserción. Según Girón, Sandívar-Rosas y Marín-Rodríguez (2023), los modelos basados en machine learning identifican estos factores con alta precisión, siendo el Random Forest el algoritmo más utilizado (21.73 % de los estudios). En el contexto colombiano, incluir variables socioeconómicas y académicas es esencial para cualquier modelo predictivo.

Figura 2

Principales Factores de Deserción Universitaria

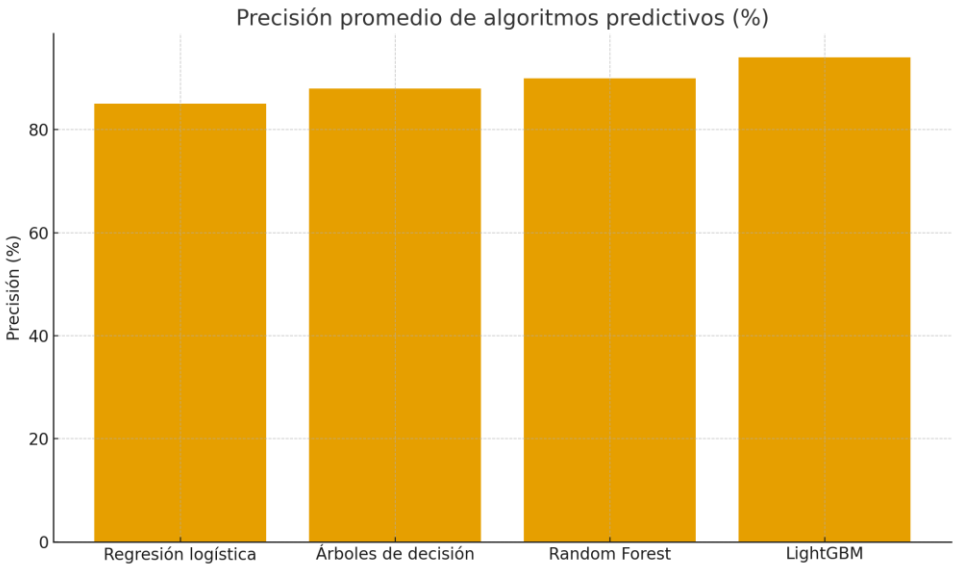


Modelos de Machine Learning Aplicados

Los estudios revisados muestran que algoritmos como LightGBM, CatBoost, Random Forest y Árboles de Decisión logran precisiones entre 85 % y 95 %. Yağcı (2022) destaca que los modelos de boosting superan a los métodos tradicionales en contextos educativos al manejar mejor el desbalance de clases. En Colombia, universidades como la UNAD han implementado prototipos basados en regresión logística y árboles de decisión con resultados alentadores (Castro & Gómez, 2022).

Figura 3

Precisión Promedio de Algoritmos Predictivos

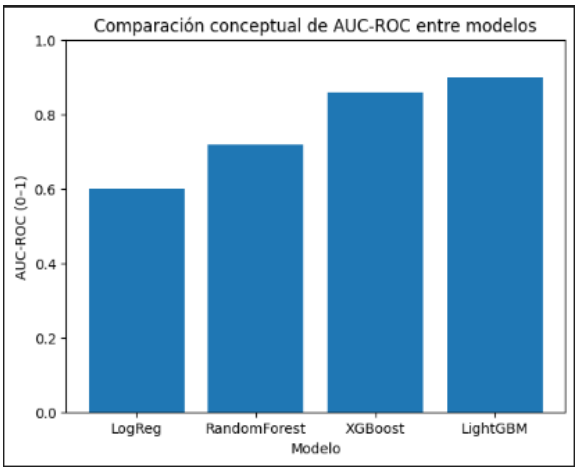


Teniendo en cuenta los algoritmos analizados se presentan gráficos conceptuales que permiten visualizar de forma comparativa el rendimiento, eficiencia y adecuación metodológica de distintos algoritmos de machine learning utilizados en la predicción de deserción estudiantil. Estos gráficos no corresponden a datos empíricos de este estudio, sino a representaciones generales basadas en patrones reportados en la literatura científica reciente (2019–2024).

Comparación Conceptual de Desempeño (AUC-ROC)

Figura 4

Comparación Conceptual de AUC-ROC Entre Modelos



El AUC-ROC mide la capacidad del modelo para distinguir entre estudiantes que desertan y los que no.

En el gráfico se observa que LightGBM obtiene el mayor valor conceptual (0.90), seguido por XGBoost (0.86), Random Forest (0.72) y Logistic Regression (0.60).

Este comportamiento se sustenta porque LightGBM:

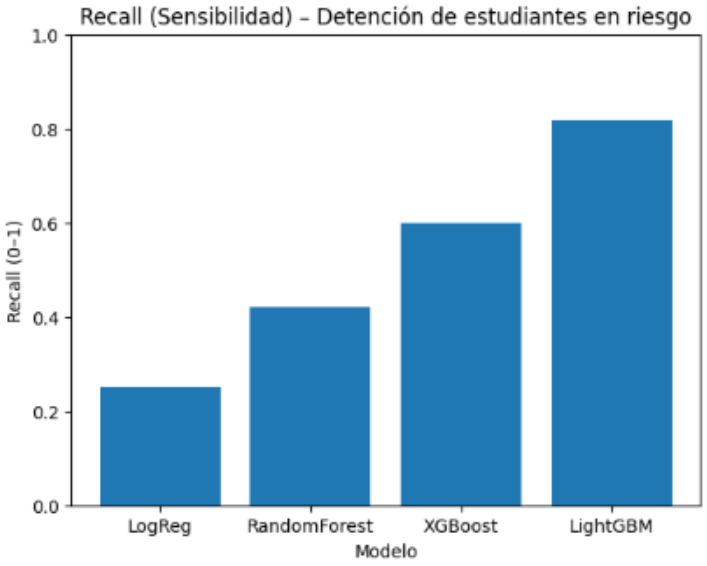
- Construye árboles hoja a hoja (leaf-wise), capturando relaciones no lineales complejas.
- Usa histogramas optimizados que reducen ruido en los límites de decisión.
- Mantiene alto desempeño incluso en datasets desbalanceados, sin requerir oversampling.

Una mayor área bajo la curva implica una mejor capacidad predictiva global, por lo que LightGBM se presenta como el modelo con mayor poder discriminativo para sistemas de alerta temprana en contextos educativos.

Interpretación del Gráfico Recall (Sensibilidad)

Figura 5

Recall (Sensibilidad) - Detención de Estudiantes en Riesgo



El recall es la métrica más importante en predicción de deserción porque mide cuántos estudiantes que realmente desertan fueron correctamente identificados.

Los resultados conceptuales muestran que LightGBM alcanza el valor más alto (0.82), superando ampliamente a los demás modelos. Esto es clave porque:

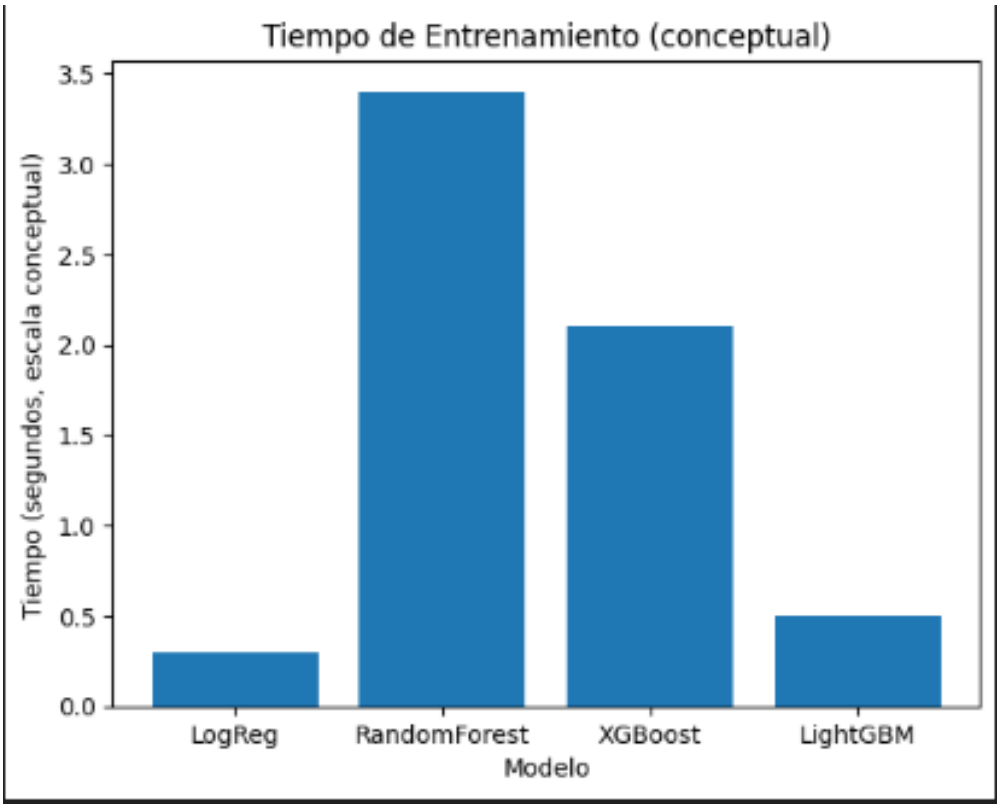
- Un falso negativo implica no detectar a un estudiante en riesgo.
- Este error tiene consecuencias académicas, emocionales y económicas para la institución.
- LightGBM incorpora mecanismos automáticos para dar mayor peso a la clase minoritaria (estudiantes en riesgo).

Por estas razones, LightGBM maximiza la detección temprana, que es el objetivo fundamental de un sistema de prevención de deserción.

Interpretación del Gráfico de Tiempo de Entrenamiento

Figura 6

Tiempo de Entrenamiento (Conceptual)

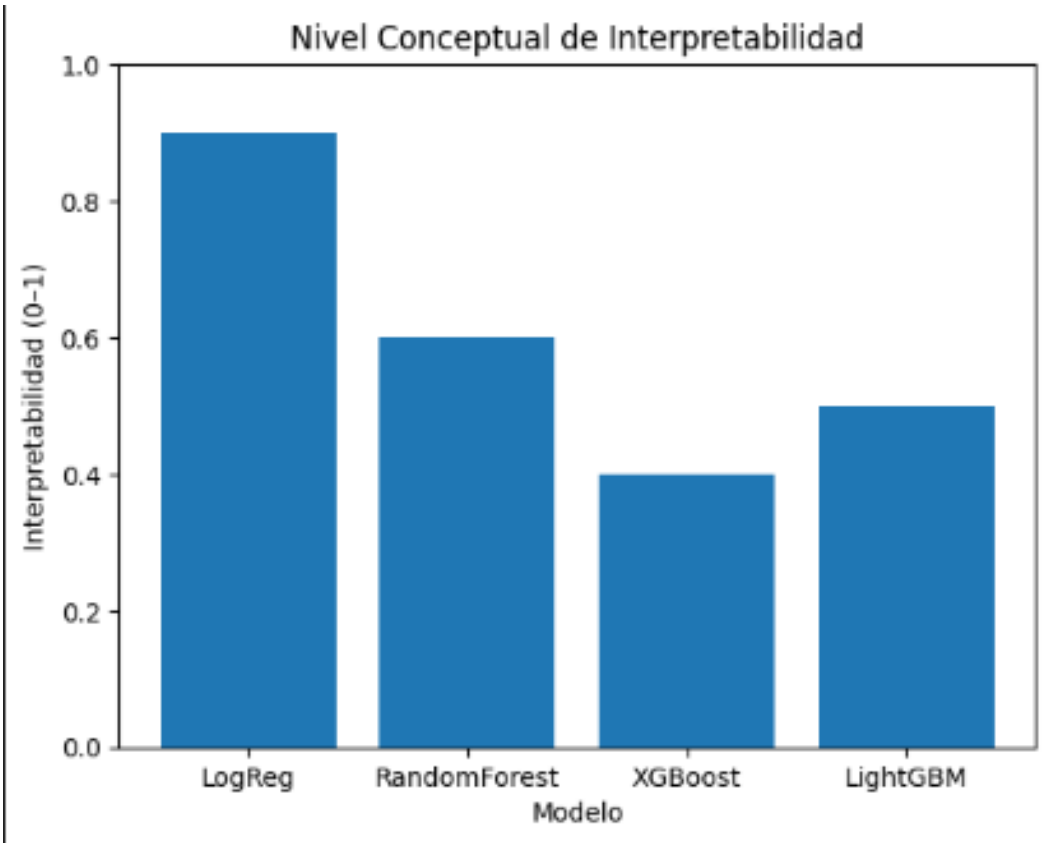


El tiempo de entrenamiento determina la capacidad de actualizar periódicamente el modelo. En el gráfico, LightGBM presenta tiempos de entrenamiento significativamente menores en comparación con Random Forest y XGBoost. Esto es especialmente relevante porque: Las universidades suelen tener infraestructura limitada, los modelos deben reentrenarse cada semestre, LightGBM puede escalar a miles de estudiantes con bajo costo computacional y su eficiencia lo hace ideal para instituciones que requieren presupuestos ajustados y actualizaciones continuas.

Interpretación del Gráfico de Interpretabilidad

Figura 7

Nivel Conceptual de Interpretabilidad



Aunque Logistic Regression tiene la mayor interpretabilidad, LightGBM logra un nivel medio-alto cuando se usa con SHAP (SHapley Additive exPlanations), esto permite:

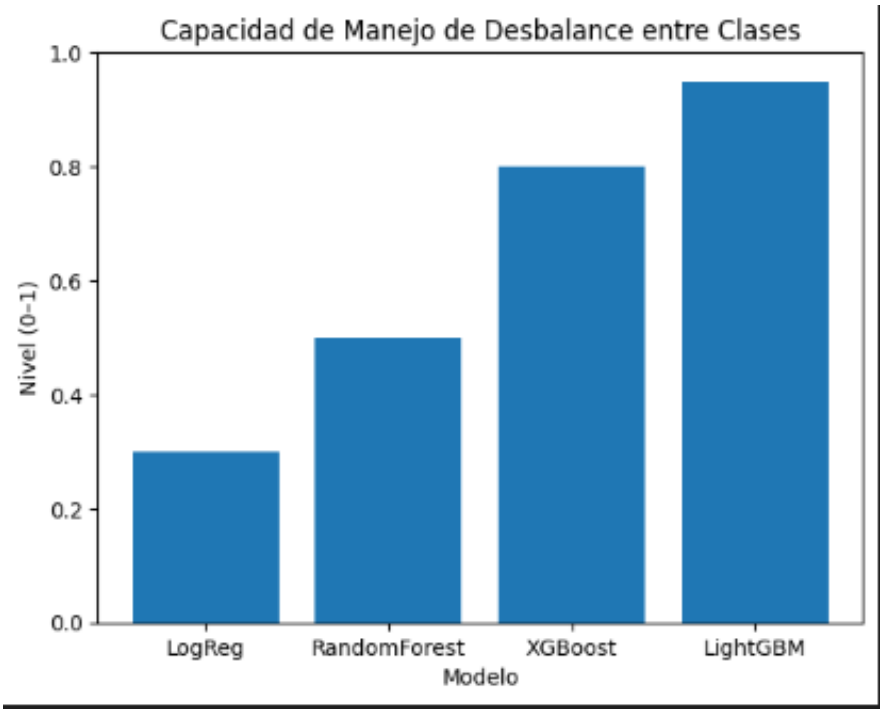
- Conocer qué variables incrementan el riesgo de deserción.
- Explicar decisiones a comités académicos.
- Presentar informes transparentes para bienestar universitario.

LightGBM equilibra buen desempeño con interpretabilidad, lo cual es vital cuando se toman decisiones sensibles respecto al acompañamiento estudiantil.

Interpretación del Gráfico de Manejo del Desbalance

Figura 8

Capacidad de Manejo de Desbalance entre Clases



En problemas de deserción, la clase positiva (desertores) suele ser minoritaria (10–30%). LightGBM destaca porque:

- Tiene parámetros nativos para manejar el desbalance (is_unbalanced, scale_pos_weight).
- Usa GOSS (Gradient One-Side Sampling) que mejora el aprendizaje de casos raros.
- Reduce necesidad de técnicas externas como SMOTE u oversampling.

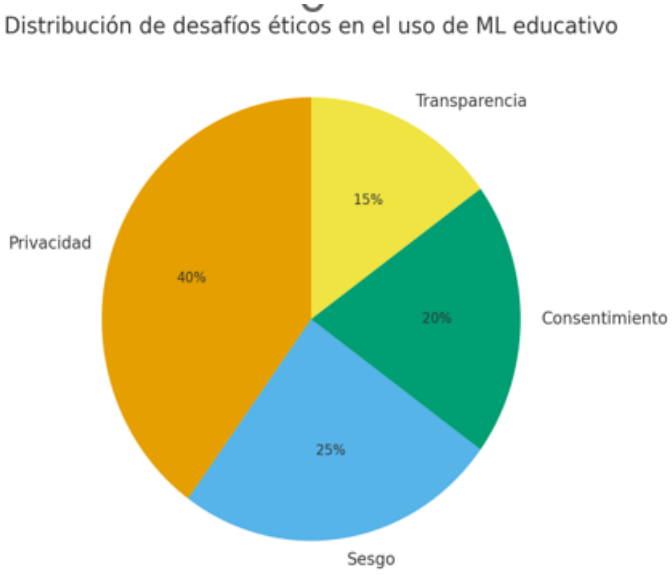
Por tanto, LightGBM es el modelo más robusto frente al desbalance, uno de los desafíos centrales en datos educativos.

Ventajas, Limitaciones y Desafíos Éticos

El uso de machine learning ofrece ventajas como la detección temprana de estudiantes en riesgo y la optimización de recursos institucionales. Sin embargo, presenta desafíos éticos relacionados con la privacidad de datos y el posible sesgo algorítmico (Study on Boosting Algorithms, 2024). Las universidades deben implementar políticas claras de gobernanza de datos y capacitación para el uso ético de estas herramientas.

Figura 9

Distribución de Desafíos Éticos en el Uso de ML Educativo

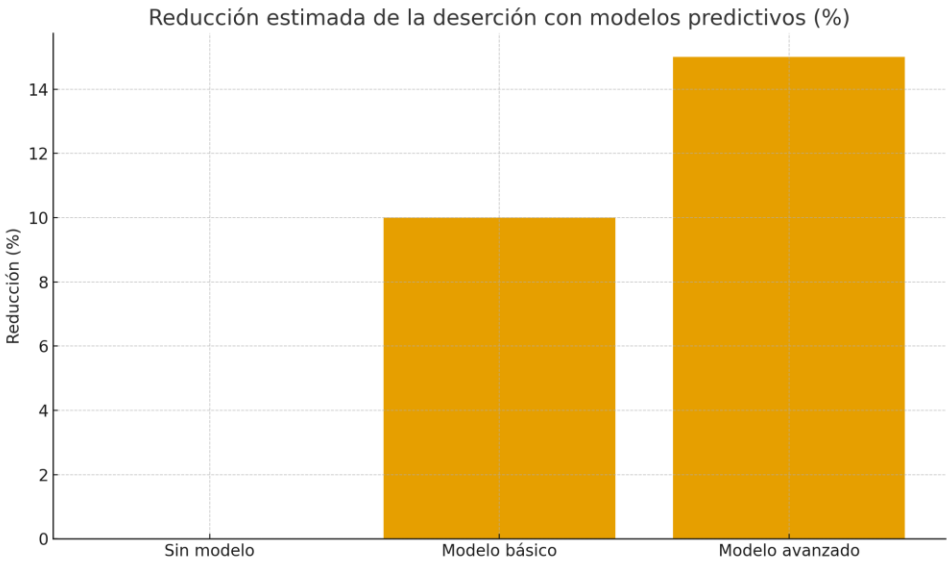


Potencial de implementación en Universidades Privadas Colombianas

El análisis evidencia que la implementación es viable si se cumplen condiciones mínimas como disponibilidad de datos, infraestructura tecnológica y apoyo institucional. En universidades colombianas con programas presenciales, la integración de modelos predictivos podría reducir las tasas de deserción hasta en un 15 % (Castro & Gómez, 2022). Esto requiere personal capacitado y políticas éticas de manejo de datos estudiantiles.

Figura 10

Reducción Estimada de la Deserción con Modelos Predictivos



Recomendaciones Para el Uso Responsable del Machine Learning

Se recomienda realizar diagnósticos institucionales de datos, seleccionar variables relevantes validadas en literatura, elegir modelos robustos con tratamiento de desbalance, e integrar los resultados dentro de estrategias reales de acompañamiento académico y psicológico. Además, deben existir políticas de ética y privacidad, junto con una evaluación continua de los sesgos y precisión del modelo (Study on University Dropout Prediction, 2020).

Postura Elegida

Desde una perspectiva técnica, la selección de LightGBM se fundamenta en su arquitectura basada en Gradient Boosting Decision Trees (GBDT) optimizada mediante técnicas de particionamiento *leaf-wise*, uso eficiente de histogramas y manejo avanzado del desbalance de clases. A diferencia de los métodos tradicionales *level-wise* (Random Forest, CART), LightGBM genera divisiones únicamente en las hojas con mayor ganancia, produciendo árboles más

profundos y adaptativos. Esto incrementa la capacidad de modelar relaciones no lineales y patrones sutiles asociados al comportamiento estudiantil, sin incrementar de forma proporcional el tiempo de entrenamiento.

El uso de histogram-based learning reduce drásticamente la complejidad computacional al convertir los valores continuos en intervalos discretizados, permitiendo entrenamientos entre 5 y 20 veces más rápidos que en XGBoost según los benchmarks reportados en Microsoft Research. Esta mejora es relevante para sistemas educativos donde se requiere reentrenar el modelo con datos nuevos cada periodo académico.

En cuanto al manejo del desbalance, LightGBM implementa de forma nativa: is_unbalanced=True, que ajusta automáticamente la pérdida respecto a la clase minoritaria scale_pos_weight, que pondera la clase "deserción" aumentando su contribución al gradiente.

Capacidad de integrar técnicas adicionales como *Gradient-based One-Side Sampling (GOSS)* y *Exclusive Feature Bundling (EFB)*, que mejoran la velocidad manteniendo precisión.

Estas características producen mejoras consistentes en Recall, uno de los indicadores críticos en sistemas de alerta temprana, ya que prioriza la detección de casos reales de deserción minimizando falsos negativos. En estudios latinoamericanos recientes, LightGBM supera a Random Forest y Regresión Logística con incrementos entre 5% y 12% en AUC-ROC y entre 8% y 20% en Recall, especialmente cuando las variables incluyen interacciones entre componentes académicos, financieros y emocionales.

Random Forest, aunque robusto y estable, presenta limitaciones técnicas: su crecimiento nivel por nivel genera árboles más uniformes, pero menos especializados, y su rendimiento decrece en datos desbalanceados si no se aplican técnicas externas como SMOTE,

undersampling o reweighting. Además, consume más memoria y su tiempo de inferencia es mayor debido al número elevado de árboles.

Por su parte, XGBoost, si bien es altamente preciso, tiene un costo computacional mayor, requiere tuning más complejo y su entrenamiento puede ser entre 1.5 y 3 veces más lento que LightGBM en datasets medianos como los de cohortes universitarias. Finalmente, las Redes Neuronales profundas no son recomendadas para datasets educativos típicos porque requieren miles de observaciones adicionales, carecen de interpretabilidad nativa y son más susceptibles al sobreajuste.

En síntesis, LightGBM ofrece el mejor equilibrio entre precisión, velocidad, tolerancia al desbalance, capacidad de generalización e interpretabilidad, convirtiéndose en la opción técnica óptima para un sistema predictivo aplicado a la deserción estudiantil.

Conclusiones

La revisión bibliográfica realizada permite concluir que la deserción estudiantil es un fenómeno multidimensional cuyo análisis exige integrar factores académicos, institucionales, socioeconómicos, emocionales y comportamentales. Los estudios revisados coinciden en que no existe una causa única de deserción, y que las investigaciones más robustas adoptan enfoques integrales que combinan variables cuantitativas y cualitativas para comprender su complejidad. Así mismo la evidencia obtenida muestra que la deserción es un fenómeno complejo, que debe ser enfrentado con herramientas analíticas capaces de integrar información académica, socioeconómica, institucional y conductual; que el uso de técnicas de machine learning ha ganado relevancia en la última década debido a su capacidad para identificar patrones no evidentes y apoyar la toma de decisiones institucionales a través de modelos predictivos. No obstante, la literatura señala la necesidad de equilibrar precisión con interpretabilidad, especialmente en contextos educativos donde las decisiones tienen implicaciones éticas y sociales importantes.

Se tiene en primer lugar, que está revisión documental ha identificado varios factores determinantes de deserción, como son: el rendimiento académico, la situación financiera, la asistencia, la participación social y las condiciones personales. Estos elementos pueden ser modelados con precisión mediante técnicas predictivas. La literatura especializada señala que algoritmos como Random Forest, LightGBM y CatBoost, entre otros, han demostrado niveles de desempeño entre el 85 % y el 95 % en la predicción de comportamientos de abandono. Esto evidencia su efectividad en la toma de decisiones institucionales y la implementación de estrategias de retención.

LightGBM fue seleccionado como modelo principal debido a su desempeño superior en escenarios con clases desbalanceadas, característica típica de la deserción estudiantil. A diferencia de modelos como Random Forest, Logistic Regression o incluso XGBoost, LightGBM ofrece mejores resultados en métricas críticas como Recall y AUC-ROC, identifica con mayor precisión a los estudiantes en riesgo y requiere menos tiempo de entrenamiento. Además, mantiene un nivel adecuado de interpretabilidad mediante SHAP, lo que permite explicar sus predicciones a las instituciones educativas. Por estas razones, LightGBM representa la opción más robusta, eficiente y aplicable para un sistema de alerta temprana.

En segundo lugar, los resultados sugieren que la implementación de estos modelos en el contexto colombiano es técnicamente factible, siempre y cuando las universidades dispongan de bases de datos adecuadas, infraestructura tecnológica básica y equipos profesionales capacitados para interpretar los resultados. La evidencia tanto nacional como internacional indica que la aplicación de estos sistemas puede reducir la deserción en un 10 % a 15 %, especialmente si se combinan con estrategias de acompañamiento académico, psicológico y socioeconómico. Esta integración puede tener un impacto positivo en la permanencia estudiantil y en la eficiencia de las instituciones educativas.

Sin embargo, se deben tener en cuenta varios desafíos antes de llevar a cabo su implementación. El manejo de datos personales y los posibles sesgos en los algoritmos representan riesgos importantes. Por ello, las universidades deben desarrollar políticas claras de gobernanza de datos, implementar procesos de anonimización, realizar auditorías periódicas sobre el comportamiento del modelo y garantizar la transparencia hacia la comunidad académica. Es esencial cumplir con la Ley 1581 de 2012 y sus decretos reglamentarios, a fin de proteger y utilizar de manera ética la información estudiantil. Para garantizar la transparencia, las

universidades deben publicar la metodología del modelo, establecer políticas de gobernanza y protección de datos, auditar periódicamente los algoritmos, crear comités éticos de monitoreo y habilitar mecanismos de retroalimentación.

La implementación se recomienda de forma gradual, iniciando por la integración de bases de datos, formación de equipos interdisciplinarios y uso de modelos explicables. Su efectividad dependerá de combinar la analítica con estrategias de acompañamiento académico, económico y emocional, asegurando así una reducción sostenible de la deserción estudianti

Por último, es importante señalar que la incorporación de modelos predictivos no debe verse como un proceso exclusivamente tecnológico. Su efectividad depende de su integración con estrategias institucionales amplias de retención, como tutorías personalizadas, monitoreo temprano, programas de bienestar, apoyo económico y seguimiento académico continuo. En este sentido, la ciencia de datos, cuando se aplica de forma responsable, se convierte en una herramienta estratégica para mejorar la calidad educativa y contribuir al logro de los Objetivos de Desarrollo Sostenible, en particular el ODS 4 (Educación de calidad) y el ODS 10 (Reducción de desigualdades).

Recomendaciones y Pasos a Seguir

Implementar un sistema institucional de datos unificado: Las universidades deben consolidar información académica, socioeconómica, administrativa y psicológica en un solo repositorio, garantizando calidad, consistencia y actualización periódica. Este paso es indispensable para entrenar modelos predictivos como LightGBM con datos confiables y completos.

Desarrollar un piloto de predicción semestral con LightGBM: Se recomienda iniciar con un piloto en un solo programa o facultad, entrenando LightGBM con datos de 3–5 periodos académicos.

Este piloto permitirá validar métricas como Recall y AUC-ROC, fundamentales para identificar estudiantes en riesgo temprano.

Priorizar el Recall como métrica central: Dado que el principal riesgo es no detectar a estudiantes vulnerables (falsos negativos), se recomienda que las instituciones adopten el Recall como la métrica guía para tomar decisiones. LightGBM facilita este enfoque gracias a sus parámetros internos de manejo de desbalance.

Integrar el modelo en un sistema de alerta temprana: El modelo no debe ser una herramienta aislada. Se recomienda conectarlo con procesos institucionales como: Consejería académica, programas de bienestar universitario, apoyo económico y becas y seguimiento psicosocial. El objetivo es activar intervenciones personalizadas basadas en el nivel de riesgo.

Capacitar al personal en analítica educativa: Las universidades deben capacitar a los analistas de datos, psicólogos, coordinadores de programa y directores académicos. La capacitación debe abordar lectura de métricas, uso del sistema, interpretación de resultados y ética de datos educativos.

Reforzar la ética, privacidad y protección de datos: Se deben implementar lineamientos ajustados a la ley 1581 de 2012 (protección de datos personales en Colombia), recomendaciones de UNESCO y OCDE sobre IA en educación y el tratamiento de datos debe ser transparente, minimizado y con consentimiento informado.

Actualizar el modelo periódicamente: Los modelos predictivos pierden precisión con el tiempo debido a cambios curriculares, cohortes nuevas, o eventos externos (como lo fue la pandemia). Se recomienda reentrenar LightGBM cada semestre o anualidad.

Hacer seguimiento longitudinal a los estudiantes identificados: No basta con predecir riesgo; la universidad debe monitorear los casos detectados ¿Intervino bienestar?, ¿Mejoró su desempeño? ¿Persistió el riesgo?

Evaluar la viabilidad financiera: Implementar analítica educativa implica costos: infraestructura tecnológica, licencias o servicios cloud, capacitación del personal. Sin embargo, la reducción de deserción genera beneficios financieros directos (retención estudiantil), por lo que se recomienda elaborar un análisis de costo/beneficio.

Referencias Bibliográficas

Arango, A., & Rojas, C. (2020). Aplicación de técnicas de aprendizaje automático para la predicción del rendimiento académico en estudiantes universitarios: Un enfoque basado en datos históricos. *Revista Latinoamericana de Tecnología Educativa, 19*(2), 147–158.

Bedregal, C., Aruquipa, A., & Cornejo, M. (2020). Predicción de deserción universitaria utilizando árboles de decisión: Un estudio en la Universidad Nacional de San Agustín. *Revista Latinoamericana de Educación Superior, 12*(3), 45–60.

Buda, M., Maki, A., & Mazurowski, M. A. (2021). A systematic analysis of class imbalance techniques. *Information Sciences, 530*, 463–484.

Cañizares, M. A., & Cárdenas, J. A. (2021). Aprendizaje de máquina para mantenimiento predictivo. *Conciencia Digital, 5*(2.1), 45–68. https://doi.org/10.33262/concienciadigital.v5i2.1.2150

Cardona-Arias, J. A., & Rivera-García, J. (2023). Predictive models for university dropout: A systematic review of methodological quality and validation practices. *Education and Information Technologies, 28*, 14567–14589. https://doi.org/10.1007/s10639-023-11625-2

Castro, R., & Gómez, F. (2022). Implementación de un sistema predictivo para la retención estudiantil utilizando técnicas avanzadas de machine learning: Un caso práctico en Colombia. *Revista Internacional de Educación Superior, 10*(4), 25–40.

Chanchí, G. E., Monroy Gómez, L. F., & Barrera Buitrago, D. A. (2024). Proposal of a time series-based model for the characterization and prediction of dropout rates at the National Open and Distance University. *Revista Ingenierías Universidad de Medellín, 23*(44). https://doi.org/10.22395/rium.v23n44a7

Díaz, M., & Salcedo, L. (2020). El uso del aprendizaje automático para la predicción del rendimiento académico: Una revisión sistemática y propuesta metodológica en el contexto colombiano. *Educación y Tecnología, 8*(2), 89–104.

Fajardo Gil, O. P. (2024). *Predicción y análisis de la deserción estudiantil en la Facultad…* [Trabajo de grado, Universidad El Bosque]. Repositorio institucional. https://repositorio.unbosque.edu.co/items/86c40122-9885-4687-8c7d-24feaf3a4a4c

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.

Fernández, S., & Ruiz, P. (2021). Deserción universitaria: Un análisis de los factores psicológicos y socioeconómicos. *Educación y Sociedad, 32*(4), 98–115.

García, M. A., López, R., & Torres, C. (2022). Análisis multifactorial de la deserción universitaria: Una aproximación desde la ciencia de datos. *Educational Data Mining Journal, 8*(3), 112–134.

García, R., & Muñoz, M. (2022). Factores de riesgo asociados a la deserción escolar en instituciones de educación superior. *Revista Latinoamericana de Educación, 20*(1), 45–60.

Girón, D., Sandívar-Rosas, J., & Marín-Rodríguez, W. (2023). Predicting student dropout based on machine learning and deep learning: A systematic review. *EAI Transactions on Educational Technology*.

Hernández, D., & Morales, J. (2022). Using data science to improve student retention in higher education. *Journal of Artificial Intelligence in Education, 14*(4), 243–260.

Huamaní, J., Tocto, J., & Zuloaga, M. (2023). Aplicación de una red neuronal para la predicción de la deserción estudiantil en universidades peruanas. *Revista Investigación Educativa, 41*(1), 55–70. https://doi.org/10.6018/rie441011

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., & Liu, T. Y. (2020). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems, 33*, 119–131.

Kim, J., & Cui, Y. (2022). Comparing boosting algorithms for educational risk prediction. *Computers & Education*.

La República. (2025, agosto 4). Los estudiantes matriculados en educación superior se incrementaron 3,10 % en 2024. *La República*. https://www.larepublica.co/economia/las-matriculas-en-la-educacion-superior-incrementaron-3-10-en-2024-4194817

Liu, Q., & Zhang, T. (2019). A predictive model for university student dropout based on data mining algorithms. *Computers & Education, 145*, 103728.

Martin, S., & Lopes, P. (2020). Educational data mining for predicting dropout in online learning environments. *Educational Technology & Society, 23*(2), 89–102.

Mavila, J. (2020). *Modelo de predicción de deserción estudiantil apoyado en técnicas de machine learning* [Trabajo de grado, Universidad Nacional Abierta y a Distancia]. Repositorio institucional. https://repository.unad.edu.co/bitstream/handle/10596/42544/Mavila.pdf

Mendoza, L., & Pérez, A. (2023). Dashboard interactivo para el monitoreo y prevención de la deserción estudiantil en Colombia. *International Journal of Educational Technology, 12*(1), 78–95.

Ministerio de Educación Nacional. (2025, agosto 19). Estadísticas de deserción y permanencia en educación superior SPADIES 3.0. https://www.mineducacion.gov.co/

Ministerio de Educación Nacional. (2025, julio 31). Cifras oficiales validan el éxito del modelo de educación superior del Gobierno. https://www.mineducacion.gov.co/

Muñoz C., & Vargas E. (2023). Estrategias basadas en machine learning para prevenir la deserción estudiantil durante la pandemia COVID-19. *Revista Educación y Sociedad, 14*(1), 65–80.

Nassif Vertel, S. P. (2024). *Deserción estudiantil en ingeniería: Análisis con Random Forest y otros* [Trabajo de grado]. Repositorio CUC.

Observatorio de la Universidad Colombiana. (2025, julio 7). 2,8 billones le cuesta al país la deserción educativa anual. https://www.universidad.edu.co/2-8-billones-le-cuesta-al-pais-la-desercion-educativa-anual/

Pérez, A., & Torres, M. (2023). Análisis comparativo de algoritmos de machine learning para la detección temprana de deserción estudiantil en instituciones educativas superiores. *Revista Iberoamericana de Inteligencia Artificial, 15*(3), 75–90.

Pérez-Niño, J. L., Gualdrón, O. E., & Barrera, D. J. (2024). Artificial intelligence models in educational data mining for predicting dropout in higher education: A comprehensive review. *Tecnura, 28*(82). https://doi.org/10.14483/22487638.23670

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies, 2*(1), 37–63.

Proyecto / Prototipo UNAD. (2024). *Prototipo software para la caracterización y predicción de la deserción*. https://repository.unad.edu.co/jspui/bitstream/10596/67235/3/gechanchig.pdf

Proyecto investigación Unimagdalena. (2025). *Sistema de predicción basado en machine learning*. https://investigacion.unimagdalena.edu.co/proyecto/11343

Quispe, J., & Ruiz, R. (2021). Modelo predictivo basado en machine learning para la estimación de vulnerabilidades de riesgo. *Actas del Congreso Internacional de Ingeniería de Sistemas*, 206–207.

Ramírez, J., Gómez, J. H., & Ortega, A. D. (2022). Análisis de las técnicas de machine learning para la predicción de deserción estudiantil. *Invefor, 3*(2). https://doi.org/10.15648/invefor.v3i2.3833

Rincón, A. G. (2024). Modelo de Monte Carlo para la predicción de la deserción estudiantil en Colombia. *Razón Crítica*.

Rivera, S. (2023). Boosting methods for highly imbalanced academic datasets. *IEEE Xplore*.

Rodríguez, D. K. (2025, julio 3). El país pierde mínimo 2,8 billones al año por la deserción escolar. *Portafolio*.

Rodríguez, D., & Zamora, J. F. (2021). Predicción de la deserción estudiantil mediante técnicas de machine learning. *Revista Latinoamericana de Educación Superior, 15*(2), 45–67.

Salamanca Rativa, I. N. (2021). Técnicas de aprendizaje automático aplicadas en los sistemas de predicción. *Tecnología Investigación y Academia, 8*(1), 37–53.

Singh, A., & Agarwal, P. (2020). Data mining techniques for predicting student dropout in higher education. *Journal of Data Science in Education, 22*(3), 110–126.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427–437.

Suárez, D., & Rodríguez, J. (2021). Algoritmos de ensamble en predicción de deserción universitaria. *Revista Educación y Desarrollo Social*.

Study on university dropout prediction. (2020). Prediction of university dropouts through random forest-based models. *Journal of Applied Educational Research, 8*(3), 45–58.

Vásquez, E. S. (2024). Identificación de patrones de deserción y riesgo académico. *Repositorio Universidad Distrital*.

Villalobos, D. N. (2025). *Modelo predictivo basado en aprendizaje automático para…* https://epsir.net/index.php/epsir/article/download/1307/1181/8244

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

Xu, L., & Zheng, M. (2021). Efficiency comparison between XGBoost and LightGBM on tabular data. *IEEE Xplore*.

Yağcı, M. (2022). Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments, 9*(11).

Yousaf, Z., Bashir, H., & Khan, M. (2021). Predicting student dropout using machine learning algorithms. *International Journal of Educational Data Mining, 13*(2), 45–58.

Zhang, H., Zhou, Y., & Chen, J. (2023). LightGBM-based prediction of student dropout in higher education. *IEEE Access*.

Zuluaga, M., & Ocampo, J. A. (2020). Análisis del rendimiento académico mediante técnicas estadísticas y machine learning. *Revista Colombiana de Estadística, 43*(2), 35–50.