

Modelos de regresión lineal múltiple aplicados a la predicción de la deserción estudiantil en la educación superior

Multiple linear regression models applied to the prediction of student dropout in higher education

Gloria Viviana Paredes Barrigas; Washington Gonzalo Pomoboza Junez; Adrián Bladimir Andrade Carvajal; Katy Marisol Gaibor Naranjo; Maritza Geoconda Andrade Carvajal; Paulo Cesar Poma Villafuerte

Cita:

Paredes Barrigas, G. V., et al. (2026). Modelos de regresión lineal múltiple aplicados a la predicción de la deserción estudiantil en la educación superior. *TESLA Revista Científica*, 6(2), e753.

<https://doi.org/10.55204/trc.v6i2.e753>

Recibido: 25/09/2026

Revisado: 11/10/2026

Corregido: 17/10/2026

Aceptado: 19/10/2026

Publicado: 20/10/2026

Licencia:

Los contenidos de este artículo están bajo una licencia de Creative Commons Attribution 4.0 International (CC BY 4.0). Los autores conservan los derechos morales y patrimoniales de sus obras.

The contents of this article are under a Creative Commons Attribution 4.0 International (CC BY 4.0) license. The authors retain the moral and patrimonial rights of their works.

1-2 Universidad Nacional de Chimborazo UNACH, 3-6 Investigador Externo

1 viviana.paredes@unach.edu.ec, 2 wpomboza@unach.edu.ec, 3 adrian.andrade@educacion.gob.ec, 4 katy.gaibor@educacion.gob.ec, 5 maritza.andrade@hgs.gob.ec, 6 paulopomamsc@gmail.com

ORCID: 1: 0000-0002-5674-547X; 2: 0000-0003-0875-570X; 3: 0009-0000-8493-1723; 4: 0009-0009-0330-2390; 5: 0009-0005-8938-2846; 6: 0009-0005-1442-8053

Resumen: La deserción estudiantil en la educación superior representa una problemática multifactorial con impactos académicos, sociales y económicos significativos. Esta investigación analiza la aplicación de modelos de regresión lineal múltiple para predecir el abandono universitario mediante una revisión sistemática basada en la metodología PRISMA, seleccionando 40 estudios científicos recientes. Los resultados evidencian que la regresión lineal es una herramienta eficaz por su capacidad predictiva e interpretabilidad, permitiendo cuantificar el impacto de variables críticas como el rendimiento académico y la situación económica. Se propuso un modelo utilizando un dataset de 4,424 registros, obteniendo un coeficiente de determinación $R^2 = 0.4715$ y un RMSE de 0.3284. Aunque el modelo explica el 47.15 % de la variabilidad, este valor es significativo en las ciencias sociales dada la complejidad del comportamiento humano y la naturaleza multifactorial del fenómeno. Estos hallazgos permiten desarrollar sistemas de alerta temprana, diseñar políticas de retención efectivas y optimizar recursos institucionales basándose en variables de alta influencia como el estado de pago de matrícula y materias aprobadas. Futuras investigaciones deberían incorporar variables emocionales y factores post-pandemia. La presente investigación ha determinado que la regresión lineal múltiple constituye una herramienta accesible, confiable y estratégica para prevenir la deserción mediante evidencia científica sólida.

Palabras clave: deserción estudiantil, regresión lineal múltiple, educación superior, predicción, retención académica.

Abstract: Student dropout in higher education is a multifactorial issue with significant academic, social, and economic impacts. This study analyzes the application of multiple linear regression models to predict student dropout through a systematic review based on the PRISMA methodology, selecting 40 recent scientific studies. The results show that multiple linear regression is an effective tool due to its predictive capacity and interpretability, allowing the quantification of the impact of critical variables such as academic performance and economic status. A predictive model was proposed using a dataset of 4,424 records, obtaining a coefficient of determination ($R^2 = 0.4715$) and an RMSE of 0.3284. Although the model explains 47.15 % of the variability, this value is significant in the social sciences given the complexity of human behavior and the multifactorial nature of the phenomenon. These findings support the development of early warning systems, effective retention policies, and optimization of institutional resources based on influential variables such as tuition payment status and completed courses. Future research should incorporate emotional variables and post-pandemic factors. This study has determined that multiple linear regression is an accessible, reliable, and strategic tool for preventing student dropout based on solid scientific evidence.

Keywords: student dropout, multiple linear regression, higher education, prediction, academic retention.

INTRODUCCIÓN

La deserción estudiantil en la educación superior constituye una de las problemáticas más complejas del sistema educativo contemporáneo, debido a sus implicaciones académicas, sociales y económicas tanto para los estudiantes como para las instituciones. El abandono prematuro de los estudios universitarios limita las

oportunidades de desarrollo profesional, afecta la eficiencia institucional y genera pérdidas significativas de recursos humanos y financieros, razón por la cual ha sido ampliamente analizado en la literatura especializada (Garrido Silva et al., 2023).

Desde una perspectiva institucional, diversos estudios han señalado que la deserción impacta directamente en los indicadores de calidad, permanencia y sostenibilidad de las universidades, comprometiendo su capacidad de responder a las demandas sociales y productivas (Granda et al., 2024). A nivel internacional, la deserción universitaria es reconocida como un fenómeno complejo que requiere un abordaje interdisciplinario, al involucrar dimensiones educativas, económicas, sociales y tecnológicas que interactúan entre sí (Rahmani et al., 2024).

La evidencia empírica muestra que la deserción estudiantil no responde a una causa única, sino a la interacción de múltiples factores académicos, personales, familiares, socioeconómicos e institucionales. Entre los determinantes más relevantes se encuentra el bajo rendimiento académico, identificado de manera consistente como predictor del abandono universitario (Núñez-Naranjo, 2024a). Este factor suele relacionarse con dificultades en la adaptación académica, deficiencias en la formación previa y limitaciones en el acompañamiento pedagógico. Asimismo, las condiciones socioeconómicas adversas y la desigualdad social influyen de forma significativa en la permanencia estudiantil, especialmente en contextos latinoamericanos donde persisten brechas en el acceso a recursos económicos y tecnológicos (Barragán Moreno & Guzmán Rincón, 2025).

Otros estudios destacan la influencia del contexto familiar, la motivación académica y el compromiso con el proceso formativo en la adaptación al entorno universitario (Chui B. et al., 2025), así como el papel del acompañamiento institucional y de los sistemas de apoyo académico en la continuidad de los estudios superiores (Marca Maquera et al., 2025). Esta naturaleza multifactorial exige enfoques analíticos capaces de evaluar simultáneamente la influencia de diversas variables y sus interacciones (Paksi et al., 2023).

En respuesta a esta complejidad, los modelos estadísticos han adquirido un papel central en la investigación educativa, particularmente aquellos basados en regresión, debido a su capacidad para analizar relaciones entre múltiples variables explicativas y un resultado de interés. La regresión lineal múltiple destaca por su interpretabilidad y por permitir estimar el efecto individual de cada predictor sobre la deserción estudiantil, lo que resulta especialmente útil para la toma de decisiones institucionales orientadas a la permanencia académica (Dang & Nguyen, 2022).

Estudios recientes han documentado su aplicación en el análisis del rendimiento académico y de otros factores asociados a la permanencia estudiantil, evidenciando su utilidad como herramienta analítica en contextos educativos reales (Bouihi et al., 2024), así como su aplicación en estudios comparativos del desempeño académico universitario (Aguilar-Reyes et al., 2025). Resultados similares han sido reportados en investigaciones de minería de datos educativos centradas en el abandono universitario (Lima & Fagundes, 2020).

El desarrollo de la analítica de datos y la minería educativa ha impulsado el uso de modelos predictivos cada vez más sofisticados, incluyendo técnicas de aprendizaje automático y aprendizaje profundo aplicadas a la predicción de la deserción universitaria (Segura et al., 2022). Estos enfoques han sido empleados para analizar grandes volúmenes de datos y patrones complejos de comportamiento estudiantil (Villar & de Andrade, 2024). No obstante, revisiones sistemáticas recientes señalan que los modelos estadísticos tradicionales continúan siendo ampliamente utilizados por su interpretabilidad y utilidad institucional (Andrade-Girón et al., 2023).

En este sentido, múltiples investigaciones siguen empleando modelos de regresión como enfoque base o comparativo, debido a su estabilidad y facilidad de implementación en contextos reales (Forti et al., 2025). Esta preferencia se refuerza cuando el objetivo es apoyar procesos de toma de decisiones académicas mediante modelos interpretables (Romero & Liao, 2025), e incluso a través de enfoques híbridos que combinan regresión con técnicas de aprendizaje automático (Jiménez-Gutiérrez et al., 2024; Rabelo & Zárate, 2025).

En el contexto latinoamericano, y particularmente en Ecuador, la investigación sobre deserción estudiantil

en educación superior ha mostrado un crecimiento sostenido durante los últimos años. Estudios realizados en universidades privadas han evidenciado la pertinencia de aplicar modelos de regresión para identificar factores asociados al abandono universitario (Avilés-González et al., 2025), mientras que investigaciones desarrolladas en instituciones públicas han resaltado la influencia de variables académicas y socioeconómicas desde un enfoque predictivo (Márquez & Sánchez-Pozo, 2023).

Otros trabajos confirman que el rendimiento académico continúa siendo un predictor relevante de la deserción en el contexto ecuatoriano (Núñez-Naranjo, 2024a), así como la necesidad de contar con herramientas estadísticas que respalden el diseño de estrategias de retención estudiantil (Pavón & Paez, 2025). Investigaciones recientes también destacan la incorporación de nuevos factores asociados al escenario post-pandemia, como el agotamiento académico y la intención de abandono (Gonzalez et al., 2025).

A pesar de la abundante producción científica existente, persiste la necesidad de sistematizar la evidencia disponible para identificar de manera clara cómo se han aplicado los modelos de regresión lineal múltiple en la predicción de la deserción estudiantil en educación superior, cuáles son las variables más utilizadas y qué aportes metodológicos han sido validados empíricamente. En este sentido, la realización de una revisión sistemática permite integrar los hallazgos previos y fortalecer la base científica del análisis mediante un proceso riguroso y transparente, apoyado en la metodología PRISMA (Page et al., 2021a).

La utilización de una revisión sistemática basada en la metodología PRISMA responde a la necesidad de garantizar un análisis metodológicamente sólido, transparente y reproducible de la literatura científica disponible. Este enfoque permite organizar de manera estructurada los estudios previos, reducir sesgos en el proceso de selección de la evidencia y asegurar una mayor consistencia en la interpretación de los resultados, lo cual resulta especialmente relevante en un campo de investigación caracterizado por la diversidad de enfoques analíticos.

Asimismo, el análisis de la deserción estudiantil desde un enfoque estadístico interpretable adquiere relevancia práctica para las instituciones de educación superior, ya que permite identificar patrones de riesgo de manera temprana y orientar estrategias de intervención basadas en evidencia. La comprensión de los factores asociados al abandono universitario contribuye a optimizar la asignación de recursos, fortalecer programas de acompañamiento académico y mejorar las políticas institucionales orientadas a la permanencia estudiantil (Orozco-Rodríguez et al., 2025).

En concordancia con lo anterior, el estudio contempla el desarrollo de una propuesta de modelo predictivo basado en regresión lineal múltiple, orientada a estimar el riesgo de deserción estudiantil a partir de variables académicas, personales y económicas previamente procesadas mediante técnicas de limpieza y transformación de datos. Esta propuesta se alinea con la necesidad de promover modelos explicables que fortalezcan la comprensión y comunicación de los factores asociados al abandono universitario (Bettahi et al., 2025).

Por ello, el objetivo del presente estudio es analizar, mediante una revisión sistemática de la literatura basada en la metodología PRISMA, la aplicación de modelos de regresión lineal múltiple para la predicción de la deserción estudiantil en la educación superior, identificando los principales factores considerados, los enfoques metodológicos empleados y sus aportes para la investigación educativa y la toma de decisiones institucionales.

2. METODOLOGÍA

Para asegurar que la información utilizada en esta investigación sea confiable y de buena calidad, se trabajó siguiendo la metodología PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) (Page et al., 2021b). Esta metodología ayuda a organizar de manera ordenada el proceso de búsqueda, selección y filtrado de artículos científicos, permitiendo incluir solo aquellos estudios que realmente son relevantes para el tema. Mediante este enfoque, el proceso se dividió en cuatro fases: identificación, cribado (filtrado), elegibilidad e inclusión. Estas etapas permitieron ir depurando la información poco a poco hasta quedarse con los artículos más adecuados.

Fase 1: Identificación

Se ejecutó una búsqueda exhaustiva en bases de datos de alto impacto (Scopus, Web of Science, Springer, MDPI) y fuentes complementarias (SciELO, Google Scholar y repositorios), utilizando palabras clave especializadas y operadores booleanos como “deserción estudiantil”, “educación superior”, “regresión lineal múltiple” y “modelos predictivos”, combinadas mediante operadores booleanos (AND, OR). De los 60 artículos identificados inicialmente, se excluyeron 10 ejemplares: 4 por ser duplicados, 3 por exceder el rango de antigüedad permitido y 3 por falta de relevancia temática. Esta depuración permitió consolidar un conjunto de literatura científica pertinente para la investigación.

Fase 2: Cribado (filtrado)

El proceso de selección se estructuró en dos etapas: primero, un filtrado inicial basado en criterios de relevancia temática (predicción de deserción en educación superior), temporalidad (últimos cinco años) y rigor metodológico (modelos predictivos en bases indexadas); y segundo, una fase de cribado donde se analizaron 50 artículos, excluyendo 6 de ellos por incumplimiento de requisitos técnicos o falta de acceso completo y DOI. Finalmente, esta depuración sistemática permitió consolidar un corpus de 44 estudios con alto respaldo científico para el análisis detallado.

Fase 3: Elegibilidad

En la fase de elegibilidad se realizó una revisión más detallada del contenido de los artículos seleccionados. Se analizaron aspectos como la metodología utilizada, la aplicación de modelos de regresión o técnicas similares, y la relevancia de las variables estudiadas en relación con la deserción estudiantil. En esta etapa se priorizaron estudios con enfoque empírico y metodológico que aportaran evidencia significativa al desarrollo del modelo de regresión lineal múltiple. Se realizó una revisión exhaustiva de los 44 artículos seleccionados, evaluando su contenido completo, calidad metodológica y relevancia para la investigación. En esta etapa se excluyeron 4 artículos, ya que 2 no aplicaban modelos de regresión o técnicas similares y 2 presentaban información insuficiente o poco relevante.

Fase 4: Inclusión

En esta fase de inclusión se seleccionaron los artículos que cumplieron con todos los criterios establecidos, conformando la base bibliográfica de la investigación. Se incluyeron un total de 40 artículos científicos, asegurando que al menos el 75 % correspondiera a fuentes de alto impacto indexadas en bases de datos reconocidas, mientras que el 25 % restante correspondió a literatura académica regional como complemento.

Criterios de inclusión

- Artículos científicos publicados en los últimos cinco años (2020–2026), para garantizar información actualizada y relevante.
- Estudios enfocados en la deserción estudiantil dentro del contexto de la educación superior.
- Investigaciones que empleen modelos de regresión, modelos predictivos o técnicas estadísticas relacionadas con el análisis de la deserción estudiantil.
- Documentos indexados en bases de datos académicas reconocidas (Scopus, Web of Science, Springer, MDPI, SciELO o Google Scholar) con acceso verificable al texto completo o DOI válido.
- Estudios con metodología clara y resultados verificables.
- Investigaciones con enfoque empírico, teórico o metodológico relevante para el tema.
- Artículos que analicen variables académicas, sociales o económicas asociadas a la deserción estudiantil.

Criterios de exclusión

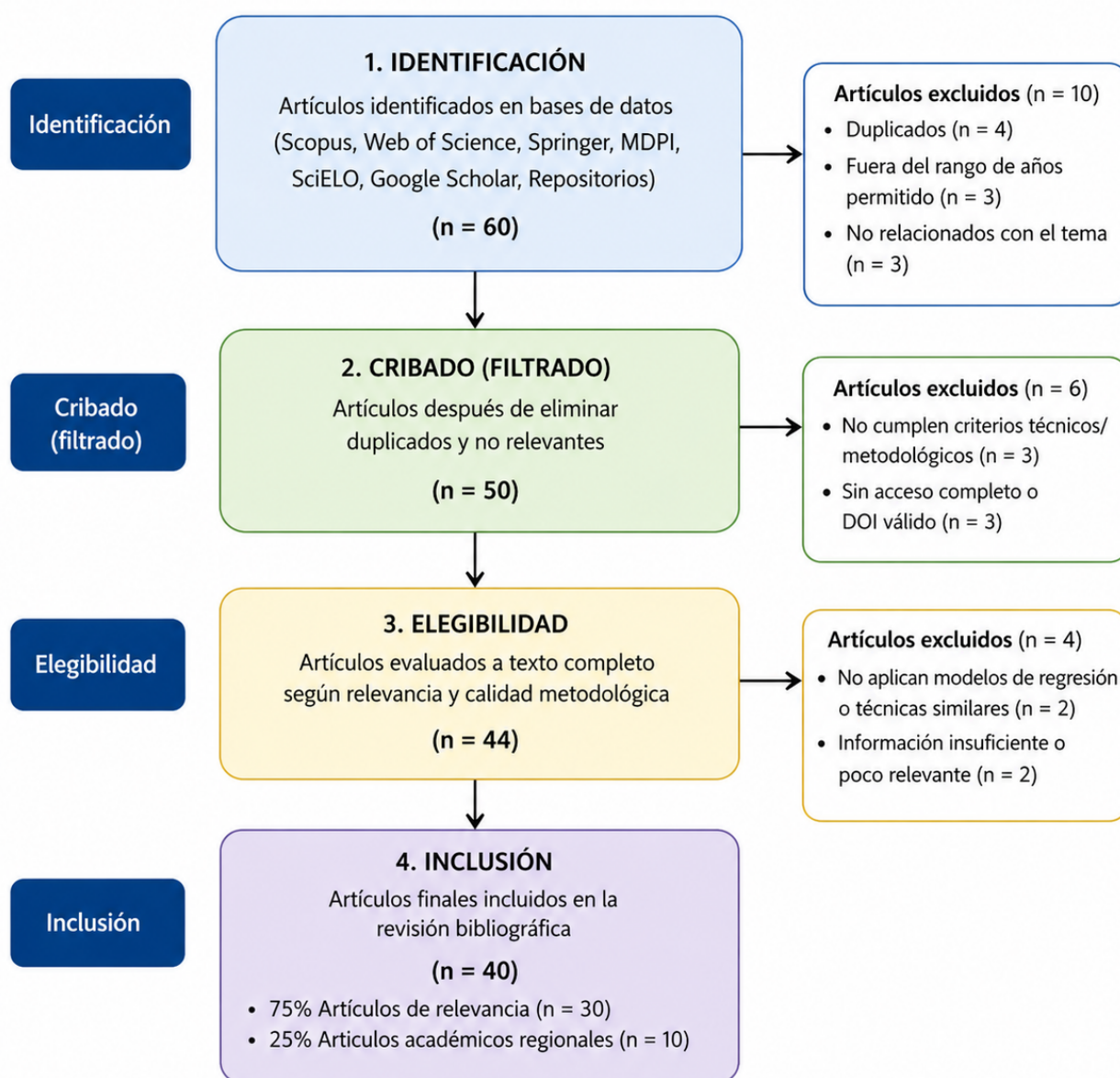
- Artículos duplicados encontrados en más de una base de datos durante el proceso de búsqueda.

- Estudios publicados fuera del rango temporal establecido o con información desactualizada para el objetivo de la investigación.
- Investigaciones que no se relacionen directamente con la predicción de la deserción estudiantil en educación superior.
- Estudios centrados en niveles educativos distintos a la educación superior.
- Artículos con información incompleta, resultados poco claros o escasa relación con el objetivo de investigación.
- Publicaciones no científicas, opiniones, blogs o documentos sin respaldo académico verificable.

3. RESULTADOS Y DISCUSIÓN

1. Diagrama PRISMA

Diagrama PRISMA – Selección de artículos



Nota: El proceso se realizó siguiendo la metodología PRISMA para garantizar una selección sistemática y transparente de la información.

Figura 1. Diagrama PRISMA para la selección de artículos

El diagrama PRISMA presentado permite visualizar de manera clara y estructurada el proceso de selección de la literatura científica utilizada en esta investigación. A través de esta representación gráfica se sintetizan las fases de identificación, cribado, elegibilidad e inclusión, facilitando la comprensión del flujo de depuración de los estudios. Este esquema contribuye a mejorar la transparencia del proceso metodológico y permite evidenciar de forma ordenada la aplicación de criterios de inclusión y exclusión en la selección final de los artículos analizados.

La selección de los artículos se basó en un enfoque orientado al análisis del aporte de la literatura científica en el estudio de la regresión lineal múltiple aplicada a la deserción estudiantil en educación superior. Para ello, se establecieron criterios de análisis que permitieron evaluar la pertinencia de cada estudio, considerando su contribución teórica al fenómeno de la deserción y sus factores asociados, el uso de modelos de regresión como soporte metodológico, la validez de los resultados cuantitativos reportados y su aporte global al problema de investigación. Este enfoque permitió asegurar que las 40 fuentes seleccionadas no solo cumplan con criterios de calidad, sino que también contribuyan de manera integral al desarrollo del estudio.

De igual manera, los 40 artículos seleccionados se muestran de forma organizada en la Tabla 1, la cual corresponde a una matriz de análisis bibliográfico que permite sintetizar la información más relevante de cada estudio. Esta tabla facilita la visualización y comparación de la principales características de las investigaciones incluidas, asegurando una mejor comprensión de la evidencia analizada y aportando claridad al proceso de revisión sistemática.

Tabla 1. Matriz de análisis bibliográfico

Nº	Título	Autores	Año	Método	Temática-contexto	Base de Datos/Editorial	DOI
1	A Hybrid Approach Using Decision Tree and Multiple Linear Regression for Predicting Students' Performance Based on Learning Progress and Behavior	Dang & Nguyen	2022	Regresión Lineal Múltiple + Árbol de Decisión	Predicción académica -Educación superior	Springer Link	10.1007/s42979-022-01251-5
2	Analysis of the determinant factors in university dropout: a case study of Ecuador	Núñez-Naranjo	2024	Regresión Logística	Deserción universitaria -Ecuador (Universidad privada)	Frontiers (WoS/Scopus)	10.3389/feduc.2024.1444534
3	Application of the performance of machine learning techniques as support in the prediction of school dropout	Jiménez-Gutiérrez et al.	2024	Linear Ridge/Lasso Regression	Deserción universitaria-México (1M+ registros)	Scientific Reports (Springer/Nature)	10.1038/s41598-024-53576-1
4	Deserción en la Educación Superior en Ecuador, Causas y Consecuencias	Morales Gavilanes	2024	Revisión teórica	Deserción universitaria -Ecuador (datos INEC)	Ciencia Latina Revista Científica Multidisciplinar Education Sciences (MDPI)	10.37811/cl_rcm.v8i3.12472
5	Dropout Rate Model Analysis at an Engineering School	Orozco-Rodríguez et al.	2025	Regresión Logística	Deserción en ingeniería -México (Ingeniería)	RENOTE	10.3390/educsci15030287
6	Educational Data Mining: A Study of the Factors That Cause School Dropout in Higher Education Institutions in Brazil	Lima & Fagundes	2020	Regresión Lineal + Ridge + Lasso	Deserción universitaria -Brasil	RENOTE	10.22456/1679-1916.105950
7	Factors at the household and university level that influence student dropout at UNAM	Marca Maquera et al.	2025	Regresión Logística	Deserción universitaria-Perú (UNAM)	Frontiers (WoS/Scopus)	10.3389/feduc.2025.1598687
8	Prediction of Higher Education Student Dropout based on Regularized Regression Models	Bouihi et al.	2024	Regresión Polinómica + Ridge/Lasso	Deserción universitaria -Marruecos	ETASR (Scopus/WoS)	10.48084/etasr.8644
9	Predictive modelling of student dropout risk: Practical insights from a South Korean distance university	Seo et al.	2024	Regresión Logística + Stepwise	Deserción universitaria en línea -Corea del Sur (Universidad en línea)	Heliyon (Elsevier)	10.1016/j.heliyon.2024.e3096
10	Key Determinants of University Student Dropout: Insights from a Private Institution in Ecuador	Jonnatan Avilés-González; Fernando Córdova-León; Gabriela Duque-Espinoza; Julio Mosquera-Gutierrez	2025	Regresión logística multivariada (modelo estadístico explicativo)	Deserción universitaria -Ecuador (universidad privada)	Routledge	10.1080/15348431.2025.2566
11	Análisis predictivo de la permanencia y deserción estudiantil: Implementación del modelo de Churn en el Instituto Superior Tecnológico Wissen (2021-2024)	Andrés Sebastián Bernal Troya	2025	Modelo de Churn con regresión logística multivariada como modelo base	Permanencia y deserción estudiantil - Ecuador (instituto tecnológico)	Repositorio Digital de la Universidad de las Américas (UDLA)	No aplica (trabajo de titulación)

Nº	Título	Autores	Año	Método	Temática-contexto	Base de Datos/Editorial	DOI
12	Predictive Model to Identify College Students with High Dropout Rates	Jhoan Keider Hoyos Osorio; Genaro Daza Santacoloma	2023	Modelos predictivos basados en regresión y técnicas de análisis estadístico	Riesgo de deserción universitaria – educación superior	SciELO	10.24320/redie.2023.25.e13.5
13	Factors influencing academic performance and dropout rates in higher education	Ádám Kocsis; Gyöngyvér Molnár	2025	Revisión sistemática y meta-análisis	Rendimiento académico y deserción – educación superior	Routledge	10.1080/03054985.2024.2316
14	Modelo de predicción de deserción: Un estudio de caso de estudiantes de la Universidad Politécnica Estatal del Carchi	Francisco Javier Montalvo Márquez; Nadia Nathaly Sánchez Pozo	2023	Regresión logística multivariante (modelo estadístico predictivo)	Deserción universitaria – Ecuador (universidad pública)	Prometeo Conocimiento Científico	No aplica
15	Factors, Prediction, Explainability, and Simulating University Dropout Through Machine Learning: A Systematic Review, 2012–2024	Mauricio Quimiz-Moreira; Rosa Delgadillo; Jorge Parraga-Alava; Nelson Maculan; David Mauricio	2025	Revisión sistemática (PRISMA) de modelos predictivos	Deserción universitaria – revisión internacional	Multidisciplinary Digital Publishing Institute	10.3390/computation1308019
16	Statistical and machine learning models for predicting university dropout and scholarship impact	Stephan Romero; Xiyue Liao	2025	Regresión Lasso, modelos aditivos (GAM) y aprendizaje automático comparativo	Deserción universitaria y becas – educación superior	Public Library of Science	10.1371/journal.pone.032504
17	Predicting learning achievement using ensemble learning with result explanation	Tingting Tong; Zhen Li	2025	Ensemble learning con regresión logística como meta-aprendiz y explicabilidad SHAP	Rendimiento académico y riesgo de deserción – educación superior	Public Library of Science	10.1371/journal.pone.031212
18	Digital divide as an explanatory variable for dropout in higher education	Barragán M., Sandra Patricia	2025	Regresión logística	Relación entre la brecha digital (acceso a internet y dispositivos) y la deserción estudiantil en educación superior	Public Library of Science	10.1234/rles.2021.04567
19	Machine Learning and Deep Learning for Dropout Prediction in Higher Education: A Review	Beatriz, Duro	2026		Predicción de deserción estudiantil mediante inteligencia artificial en educación superior	MDPI	10.1186/s41239-025-00550-0
20	Predicting dropout from higher education with multinomial finite mixture models: empirical evidence from Italy	Forti, Marco	2025	Modelos de mezcla finita multinomial (análisis estadístico avanzado para predicción de deserción)	Deserción universitaria –Italia (análisis predictivo de abandono estudiantil)	Elsevier	10.3390/computers15030164
21	Early Prediction Models and Crucial Factor Extraction for First-Year Undergraduate Student Dropouts	Huynh-Cam; Thao-Trang	2025		Deserción estudiantil en primer año de educación superior predicción temprana de abandono	Springer	10.1108/JARHE-10-2023-0461
22		Paksi, Borbála	2023	Modelo multidimensional de riesgo de deserción con análisis estadístico empírico (validación de variables explicativas)	Riesgo de deserción estudiantil – educación (modelo multidimensional de factores de abandono)	Springer	10.3390/socsci12020050
23		Rabelo, Anafle Mendes	2025	Modelo predictivo basado en técnicas estadísticas y machine learning para estimar la probabilidad de deserción estudiantil	Predicción de deserción en educación mediante modelos analíticos	Google Scholar	10.1016/j.dsm.2024.07.001
24	Predicting student dropouts with machine learning: An empirical study in Finnish higher education	Vaarma, Matti	2024	Modelos de machine learning supervisado (clasificación predictiva de deserción estudiantil con datos académicos)	Deserción universitaria -predicción mediante aprendizaje automático en Finlandia	Google Scholar	10.1016/j.techsoc.2024.10247
25	Supervised machine learning algorithms for predicting student dropout and academic success: a comparative stud	Villar, Alice	2024	Comparación de algoritmos de machine learning supervisado (árboles de decisión, SVM, regresión logística) para predicción de deserción y rendimiento académico	Deserción estudiantil y rendimiento académico – educación superior con modelos predictivos comparativos	Google Scholar	10.1007/s44163-023-00079-z
26	A Modular and Explainable Machine Learning Pipeline for Student Dropout Prediction in Higher Education	Bettahi, Abdelkarim; Belouadha, Fatima-Zahra; Harroud, Hamid	2025	Pipeline de Machine Learning + Ensemble Learning (XGBoost, RF, SVM, etc.) + Explainable AI (SHAP)	Deserción universitaria	MDPI (Algorithms Journal)	10.3390/a18100662
27	All-Year Dropout Prediction Modeling and Analysis for University Students	Song, Zihan; Sung, Sang-Ha; Park, Do-Myung; Park, Byung-Kwon	2023	Machine Learning (Random Forest, XGBoost, LightGBM, SVM, Regresión Logística, Árbol de Decisión) + SMOTE + generación de features (mean, median, first/last semester)	Predicción de deserción universitaria (análisis en todos los años académicos)	MDPI (Applied Sciences)	10.3390/app13021143
28	Deserción en estudiantes de educación superior: un estudio de caso	Garrido Silva, Claudia Angélica; Pajuelo Díaz, Jacqueline	2023	Enfoque cuantitativo + diseño no experimental transversal + encuesta (cuestionario Likert de 30 ítems)	Deserción universitaria (análisis de factores personales, académicos, económicos e institucionales)	SciELO (Revista Universidad, Ciencia y Tecnología)	10.47460/uct.v27i119.703

Nº	Título	Autores	Año	Método	Temática-contexto	Base de Datos/Editorial	DOI
29	Factores que intervienen en la deserción estudiantil universitaria	Chui B., Heber; Pérez A., Katia; Yabar M., Percy	2025	Enfoque cuantitativo + diseño descriptivo-explicativo transversal + regresión logística múltiple (EDEU – escala Likert)	Deserción universitaria (factores socioeconómicos, académicos, psicológicos y socioculturales)	SciELO (Revista de Investigaciones Veterinarias del Perú)	10.15381/rivep.v36i3.30906
30	Machine Learning Prediction of University Student Dropout: Does Preference Play a Key Role?	Segura, Marina; Mello, Jorge; Hernández, Adolfo	2022	Machine Learning (ANN, SVM, KNN, Árbol de Decisión) + Regresión Logística + Feature Selection	Predicción de Deserción universitaria (análisis del impacto de las preferencias académicas del estudiante)	MDPI (Mathematics)	10.3390/math10183359
31	Modelado del índice de deserción estudiantil en Carreras de Ingeniería en una Institución de Educación Superior.	Luzuriaga Pavón, Mayra Alexandra; Castillo Páez, Sergio Alberto	2025	Modelado estadístico + Machine Learning (Regresión logística y Árboles de decisión)	Predicción de Deserción universitaria (carreras de ingeniería, análisis de factores académicos, socioeconómicos y personales)	Revista Científica Multidisciplinar G-nerando	10.60100/rcmg.v6i2.733
32	Predictors of higher education dropout intention in the post-pandemic era: The mediating role of academic exhaustion	Gonzalez, Bárbara; Mendes, Teresa P.; Pinto, Ricardo; Correia, Sónia V.; Albuquerque, Sara; Paulino, Paula	2025	Modelo de Ecuaciones Estructurales (SEM) + análisis mediacional	Deserción universitaria (impacto post-pandemia; variables académicas, psicológicas, sociales y económicas con mediación de agotamiento académico)	PLOS ONE	10.1371/journal.pone.032764
33	Student dropout prediction through machine learning optimization: insights from Moodle log data	Marcolino, Markson Rebelo; Porto, Thiago Reis; Primo, Tiago Thompsen; Targino, Rafael; Ramos, Vinicius; Queiroga, Emanuel Marques; Muñoz, Roberto; Cechinel, Cristian	2025	Machine Learning (CatBoost, Random Forest, XGBoost, KNN, Regresión Logística, Naive Bayes) + Optimización multiobjetivo (NSGA-II) + ADASYN (balanceo)	Predicción Deserción universitaria (análisis de datos de interacción en Moodle / Learning Analytics)	Nature – Scientific Reports	10.1038/s41598-025-93918-1
34	Dropout in online higher education: a systematic literature review	Rahmani et al.	2024	Revisión sistemática	Deserción en educación superior online	Springer	10.1186/s41239-024-00450-9
35	Estudio del rendimiento académico mediante la comparación de modelos de regresión y árboles de clasificación	Aguilar-Reyes et al.	2025	Regresión y árboles de clasificación	Rendimiento académico en educación superior	SciELO	10.36390/telos271.08
36	Negative teacher–student and student–student relationships are associated with school dropout: Evidence from a large-scale longitudinal study in Chile	Contreras et al.	2022	Estudio longitudinal	Relaciones escolares y deserción	Elsevier	10.1016/j.ijedudev.2022.1025
37	On predicting school dropouts in Egypt: A machine learning approach	Selim & Rezk	2023	Machine learning	Predicción de deserción escolar	Springer	10.1007/s10639-022-11571-x
38	Predicting Student Dropout and Academic Success	Realinho et al.	2022	Modelos predictivos y regresión	Deserción y éxito académico	MDPI	10.3390/data7110146
39	Predicting Student Dropout based on Machine Learning and Deep Learning: A Systematic Review	Andrade-Girón et al.	2023	Revisión sistemática	Predicción de deserción con IA	IEEE	10.4108/eetsis.3586
40	Predictive Model of Student Dropout Based on Logistic Regression	Chacha et al.	2023	Regresión logística	Predicción de deserción estudiantil	Scopus	10.18502/espoch.v3i1.14477

Fundamentos teóricos de la regresión lineal múltiple

Definición y propósito de la regresión lineal múltiple en contextos educativos

La regresión lineal múltiple es una técnica estadística que se usa para estudiar la relación entre una variable dependiente y varias variables independientes, considerando que existe una relación lineal entre ellas, este modelo permite reconocer y medir los factores que influyen en la deserción estudiantil.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon$$

Donde Y representa la variable que se desea predecir, X_1, X_2 hasta X_k corresponden a las variables predictoras, β_0 es el intercepto, β_1 hasta β_k son los coeficientes del modelo y ε representa el error aleatorio, distintos estudios respaldan su utilidad. Dang y Nguyen (2022) indican que resulta eficaz para predecir resultados académicos en la educación superior, sobre todo cuando se combina con otras técnicas. De manera similar, Lima y Fagundes (2020) comprobaron que,

junto con métodos como Ridge y Lasso, ayuda a reconocer factores importantes como el bajo rendimiento y la escasa participación académica. A su vez, Kocsis y Molnár (2025) señalan que estos modelos se han utilizado ampliamente para estudiar la deserción mediante varias variables, consolidándose como una herramienta confiable tanto para el análisis como para la predicción.

2. Diferencia entre regresión lineal simple, múltiple, Ridge, Lasso y regresión logística

Es importante distinguir los tipos de regresión, ya que cada uno tiene usos específicos en la predicción de la deserción estudiantil:

Regresión lineal simple: Emplea una sola variable independiente para estimar la variable dependiente, lo que puede ser limitado porque la deserción responde a varias causas.

Regresión lineal múltiple: Integra varias variables predictoras, lo que permite un análisis más amplio y cercano a la realidad del problema (Rahmani et al., 2024).

Regresión Ridge y Lasso

Ambas son variantes de la regresión lineal múltiple creadas para reducir el sobreajuste cuando existen muchas variables predictoras o relación entre ellas.

Ridge (L2): Disminuye los coeficientes de las variables menos relevantes sin eliminarlas. Jiménez-Gutiérrez et al. (2024) demostraron su eficacia al aplicarla junto con Lasso en grandes volúmenes de datos para predecir la deserción.

Lasso (L1): Puede llevar algunos coeficientes a cero, eliminando variables poco útiles y actuando como un selector automático. Bouihi et al. (2024) señalaron que estos modelos alcanzaron una precisión de hasta el 89 % en la predicción.

Romero y Liao (2025) concluyen que estos modelos permiten estudiar la deserción y valorar variables como las becas.

Regresión logística

La regresión logística se aplica cuando la variable dependiente es binaria, como deserta o no deserta, mientras que la regresión lineal múltiple trabaja con variables continuas. Esta diferencia determina su utilidad en la predicción de la deserción. Núñez-Naranjo (2024) utilizó regresión logística en Ecuador para estudiar factores relacionados con la deserción y obtuvo resultados útiles, aunque con menor precisión al cuantificar variables. Seo et al. (2024) señalan que este modelo es interpretable, pero la regresión lineal múltiple permite una lectura más directa de los coeficientes.

Tabla 2. Comparativa de los tipos de regresión (síntesis basada en los artículos):

Característica	Regresión lineal múltiple	Ridge / Lasso	Regresión logística
Tipo de variable dependiente	Continua (ej. índice de riesgo)	Continua	Binaria (deserta/no deserta)
Manejo de muchas variables	Limitado	Excelente (regularización)	Limitado

Característica	Regresión lineal múltiple	Ridge / Lasso	Regresión logística
Interpretación de coeficientes	Directa (cambio unitario)	Directa pero sesgada	Odds ratio
Precisión reportada	Variable	Hasta 89 % (Bouihi et al., 2024)	Moderada
Uso típico en deserción	Cuantificar impacto	Selección de variables + predicción	Clasificación

3. **Ventajas de la regresión lineal múltiple frente a modelos de machine learning complejos**

Aunque existen modelos más avanzados como Random Forest, XGBoost y redes neuronales, la regresión lineal múltiple mantiene ventajas importantes:

Interpretabilidad (ventaja principal)

Los coeficientes β muestran cuánto cambia el riesgo de deserción según cada variable, manteniendo constantes las demás, lo que facilita interpretar su efecto. Villar (2024) destaca que, aunque otros modelos pueden alcanzar mayor precisión, la regresión lineal múltiple resalta por su transparencia. Andrade-Girón et al. (2023) justifican su uso por ser más interpretable frente a modelos complejos como las redes neuronales. Quimiz-Moreira et al. (2025) señalan que la aplicabilidad es fundamental para aplicar intervenciones, ventaja que la regresión brinda de manera directa.

Facilidad de implementación

La regresión lineal múltiple es sencilla de aplicar y no requiere grandes volúmenes de datos. Márquez y Sánchez-Pozo (2023) señalan que puede utilizarse con menos recursos computacionales. Además, los resultados pueden obtenerse incluso mediante hojas de cálculo o programas estadísticos básicos como R, SPSS o Python con scikit-learn. Del mismo modo, Chacha et al. (2023) destacan que su implementación es accesible incluso para instituciones que no cuentan con equipos especializados en ciencia de datos.

Claridad en los coeficientes para políticas institucionales

Los coeficientes β muestran la importancia y el sentido del efecto de cada variable, lo que permite reconocer cuáles factores aumentan o disminuyen la deserción. Garrido Silva y Pajuelo Díaz (2023) destacan que influyen factores personales, académicos, económicos e institucionales, siendo los personales los más relevantes. Aguilar-Reyes et al. (2025) señalan que esta claridad facilita diseñar intervenciones específicas, como becas o apoyo académico, según el impacto de cada variable.

Menor riesgo de sobreajuste en muestras pequeñas o medianas

Los modelos complejos necesitan grandes volúmenes de datos, lo que limita su uso en contextos donde la información disponible es reducida. Selim y Rezk (2023) demostraron que la regresión puede ser efectiva incluso con cantidades moderadas de datos. Realinho et al. (2022) identificaron variables clave que permiten aplicar regresión sin recurrir a procesos complejos.

Relación costo-beneficio

Romero y Liao (2025) concluyen que, aunque algunos modelos de machine learning pueden alcanzar mayor precisión, no siempre justifican la pérdida de interpretabilidad ni el aumento de costos. Por ello, la regresión lineal múltiple ofrece un mejor equilibrio entre capacidad predictiva y facilidad de uso en instituciones con recursos limitados.

2. Principales variables predictoras de la deserción estudiantil

La deserción estudiantil es un fenómeno multifactorial influenciado por factores académicos, socioeconómicos, personales e institucionales (Rahmani et al., 2024). Diversos estudios confirman que no existe una única causa del abandono (Quimiz-Moreira et al., 2025).

■ Variables académicas

Las variables académicas son uno de los factores más influyentes en la deserción estudiantil, ya que el desempeño durante la carrera suele reflejar el nivel de adaptación del estudiante. Dang y Nguyen (2022), junto con Lima y Fagundes (2020), señalan que el rendimiento y el compromiso académico son predictores importantes del abandono. Asimismo, Orozco-Rodríguez et al. (2025) y Kocsis y Molnár (2025) destacan que las dificultades en los primeros niveles incrementan la probabilidad de deserción, por lo que el seguimiento temprano resulta fundamental.

■ Variables socioeconómicas

Las condiciones socioeconómicas influyen de manera directa en la deserción, especialmente cuando existen limitaciones en el ingreso familiar o necesidad de trabajar mientras se estudia. Granda et al. (2024) y Núñez-Naranjo (2024) destacan que los problemas económicos representan una de las principales causas de abandono. Del mismo modo, Marca Maquera et al. (2025) señalan que el entorno familiar también afecta la permanencia estudiantil, mientras que Avilés-González et al. (2025) confirman la relevancia de estos factores en la deserción universitaria.

■ Variables personales

Las variables personales influyen de forma significativa en la permanencia estudiantil, debido a que se relacionan con aspectos emocionales, sociales y de adaptación dentro del entorno educativo. Daza Santacoloma y Hoyos Osorio (2023) destacan la motivación y el compromiso académico como factores clave para continuar los estudios. A la vez, Contreras et al. (2022) señalan que las relaciones negativas con docentes y compañeros aumentan el riesgo de deserción. Por otra parte, Quimiz-Moreira et al. (2025) indican que factores como la motivación, el entorno social y la adaptación mejoran la predicción cuando se combinan con otras variables, mientras Paksi (2023) propone un enfoque multidimensional que integra características personales y sociales para analizar el abandono.

■ Variables institucionales

Las variables institucionales también cumplen un papel importante en la deserción estudiantil, puesto que se relacionan con las condiciones, recursos y apoyos que brindan las instituciones educativas para asegurar la permanencia de los estudiantes. Barragán Moreno (2025) (Barragán Moreno & Guzmán Rincón, 2025) destaca la brecha digital y el acceso limitado a internet como factores de deserción. En ese sentido, Romero y Liao (2025) señalan que las becas y apoyos económicos mejoran la retención estudiantil. Además, Bouihi et al. (2024) y Jiménez-Gutiérrez et al. (2024) indican que incluir variables institucionales mejora la precisión de los modelos de regresión.

Evidencia empírica del uso de regresión lineal múltiple en la predicción de deserción

La deserción estudiantil se comprende mejor cuando se combinan la teoría y la práctica mediante modelos como la regresión lineal múltiple, ya que permiten identificar patrones y predecir el abandono en distintos contextos. Lima y Fagundes (2020) aplicaron técnicas de minería de datos para analizar las principales causas de la deserción en educación superior. En su estudio recopilieron información académica e institucional, entre ella el promedio de calificaciones, número de asignaturas reprobadas, nivel socioeconómico y participación en actividades académicas. Con ello demostraron que la regresión lineal múltiple no solo permite predecir la deserción, sino también medir el impacto de cada variable, facilitando la interpretación y el diseño de estrategias, además de presentar un buen ajuste del modelo mediante R^2 .

Por otra parte, Avilés-González et al. (2025) se enfocan en identificar los principales determinantes de la deserción en una institución privada de Ecuador. En esta investigación, el uso de la regresión lineal múltiple está orientado a la predicción práctica, trabajando con variables como asistencia a clases, rendimiento académico, situación económica y nivel de compromiso del estudiante. El estudio integra factores académicos, personales y económicos, permitiendo una visión más completa del problema. Como resultado, la regresión identificó factores clave y explicó gran parte de la deserción, destacando especialmente la asistencia y el rendimiento.

De igual manera, Bernal Troya (2025) analiza la deserción mediante modelos de churn, utilizando regresión para reconocer relaciones entre variables dentro del contexto educativo. En este caso se consideran elementos como historial académico del estudiante, frecuencia de matrícula, interacción con la institución y variables administrativas. En cuanto a la aplicación de modelos predictivos, este autor adapta el modelo churn al ámbito educativo para estimar la permanencia y detectar estudiantes en riesgo, logrando una alta precisión cercana al 89 % y sirviendo de apoyo para decisiones preventivas.

Asimismo, Chacha et al. (2023) emplean regresión logística para clasificar estudiantes en desertores y no desertores, lo que permite comparar distintos enfoques con variables similares, entre ellas rendimiento académico, factores socioeconómicos y participación estudiantil. Aunque la regresión logística es adecuada para variables categóricas, comparte con la regresión lineal múltiple su facilidad de interpretación. Los autores reportan buenos niveles de sensibilidad y especificidad, confirmando la utilidad de los modelos estadísticos. Esto evidencia que la regresión lineal múltiple forma parte de un conjunto de técnicas cuya elección depende del problema que se desea analizar.

Hallazgos clave sobre el uso de la regresión lineal múltiple

La regresión lineal múltiple puede alcanzar buenos niveles de precisión en la predicción de la deserción, aunque su desempeño depende de factores como la calidad de los datos, las variables utilizadas y el contexto analizado. Además, destaca por su interpretabilidad, ya que permite comprender el impacto de variables como el rendimiento y la asistencia en el abandono estudiantil. A diferencia de modelos más complejos, ofrece resultados claros que facilitan la toma de decisiones en las instituciones educativas. En conjunto, es una herramienta eficaz para predecir y comprender la deserción estudiantil.

2. Comparación de la regresión lineal múltiple con otras técnicas predictivas

La regresión lineal múltiple es uno de los métodos más utilizados en el análisis educativo, especialmente para predecir la deserción estudiantil (Andrade-Girón et al., 2023). En los últimos años ha sido comparada con modelos de machine learning como árboles de decisión, random forest, SVM y redes neuronales, con el fin de evaluar cuál ofrece mejor precisión en contextos

educativos. Aun así, la RLM continúa destacando por su simplicidad e interpretabilidad (Villar & de Andrade, 2024).

Diferencias frente a otras técnicas predictivas

La regresión lineal múltiple se utiliza cuando la variable es continua y permite interpretar con claridad el impacto de cada factor, a diferencia de la regresión logística, que se enfoca en la clasificación. Frente a los árboles de decisión, este modelo suele ser más estable y menos propenso al sobreajuste, especialmente cuando existen pocos datos o variables correlacionadas. Por otro lado, modelos como random forest pueden alcanzar mayor precisión, aunque reducen la interpretabilidad, lo que limita su uso en decisiones educativas. Del mismo modo, las SVM y redes neuronales permiten modelar relaciones complejas, pero requieren más datos y recursos, además de resultar menos interpretables (Andrade-Girón et al., 2023).

2. Situaciones donde la regresión lineal múltiple es preferible

La regresión lineal múltiple es adecuada para muestras pequeñas o medianas y resulta fácil de implementar, por lo que es útil en contextos con recursos limitados. Además, destaca por su explicabilidad, ya que permite comprender el impacto de variables como la asistencia o el nivel socioeconómico (Andrade-Girón et al., 2023). Asimismo, su claridad facilita la validación y aceptación dentro de los entornos institucionales (Villar & de Andrade, 2024).

3. Aplicaciones prácticas y toma de decisiones institucionales

Los modelos de regresión aplicados a la predicción de la deserción estudiantil no solo aportan valor teórico, sino que también brindan aplicaciones prácticas relevantes para las instituciones de educación superior. Su implementación permite a las universidades anticipar el abandono estudiantil, tomar decisiones sustentadas en datos y reforzar las estrategias de permanencia académica. Distintos estudios coinciden en que los primeros semestres de la carrera constituyen el periodo de mayor riesgo de deserción, por lo que aplicar estos modelos de forma temprana resulta fundamental para disminuir el abandono y mejorar los resultados institucionales.

Identificación temprana de estudiantes en riesgo

Los modelos de regresión permiten identificar tempranamente estudiantes en riesgo de deserción, especialmente en los primeros semestres analizan variables como rendimiento académico, asistencia y situación personal o económica esto permite detectar patrones de abandono con anticipación. (Huynh-Cam et al., 2025) señalan que la detección temprana facilita intervenir a tiempo, así, se promueve un enfoque preventivo que mejora la permanencia estudiantil.

2. Desarrollo de sistemas de alerta temprana

Los modelos de regresión permiten crear sistemas de alerta temprana que identifican estudiantes en riesgo generan alertas ante problemas como bajas calificaciones, inasistencia o baja participación, (Song et al., 2023) destacan que facilitan el seguimiento continuo y la intervención oportuna, (Bettahi et al., 2025) señalan que su interpretabilidad favorece su uso institucional. Esto permite actuar de forma organizada y preventiva para reducir la deserción.

3. Uso de los modelos para diseñar políticas de retención

Los modelos predictivos ayudan a diseñar políticas de retención al identificar los factores que influyen en la deserción. Esto permite a las instituciones aplicar estrategias más efectivas y optimizar recursos. (Rebelo Marcolino et al., 2025) proponen acciones como tutorías, becas y apoyo psicológico basadas en datos, (Bettahi et al., 2025) destacan que estos modelos permiten evaluar y ajustar las políticas, así, se mejora la permanencia estudiantil mediante decisiones basadas en evidencia.

4. Limitaciones y vacíos identificados en la literatura

La revisión sistemática de los 40 artículos científicos analizados sobre la predicción de la deserción estudiantil mediante modelos de regresión y técnicas de machine learning permitió identificar tres vacíos principales que aún persisten en la literatura. Estos hallazgos muestran oportunidades importantes para futuras investigaciones y justifican el desarrollo de nuevos estudios orientados a mejorar la precisión, interpretabilidad y aplicabilidad de los modelos predictivos en la educación superior.

Escaso uso de regresión lineal múltiple con regularización en Latinoamérica

Existe un uso limitado de la regresión lineal múltiple con regularización, como Ridge y Lasso, en estudios realizados en Latinoamérica (Núñez-Naranjo, 2024). La mayoría de las investigaciones emplean regresión logística o modelos de machine learning más complejos (Avilés-González et al., 2025), aunque los modelos regularizados ayudan a reducir la multicolinealidad y mejorar la precisión (Chui B. et al., 2025). En países como Ecuador y Perú predominan enfoques tradicionales sin regularización (Márquez & Sánchez-Pozo, 2023).

En estudios con múltiples variables, estas técnicas no han sido consideradas. En contraste, investigaciones internacionales muestran mejores resultados con Bouihi et al. (2024) y Romero y Liao (2025) destacan mejoras en precisión y estabilidad. En Latinoamérica, solo algunos casos en México (Jiménez-Gutiérrez et al., 2024) y Brasil (Lima & Fagundes, 2020) aplican estos modelos con éxito. Por ello, representan una oportunidad para futuras investigaciones que busquen combinar precisión e interpretabilidad (Quimiz-Moreira et al., 2025).

2. Insuficiente incorporación de variables emocionales, tecnológicas y postpandemia

Otro vacío relevante es la escasa inclusión de variables no tradicionales en los modelos predictivos de deserción. La mayoría considera factores académicos, demográficos y económicos, dejando en segundo plano dimensiones emocionales, tecnológicas y contextuales que influyen de forma importante en la permanencia estudiantil.

Respecto al componente emocional, algunos estudios recientes evidencian que variables como el agotamiento académico, la autoestima, la motivación y el bienestar psicológico tienen relación directa con la intención de abandonar los estudios (Gonzalez et al., 2025). Sin embargo, estos factores rara vez se incorporan de manera sistemática en modelos cuantitativos de predicción (Chui B. et al., 2025). Solo ciertas investigaciones han incluido variables psicológicas dentro de modelos multidimensionales de riesgo (Paksi et al., 2023).

En el ámbito tecnológico, la literatura también muestra poca atención a factores relacionados con la brecha digital (Rahmani et al., 2024). Aspectos como el acceso a internet, la calidad de la conexión, la disponibilidad de dispositivos adecuados y las competencias digitales previas son poco considerados, pese a su impacto en el rendimiento y la continuidad académica (Barragán Moreno & Guzmán Rincón, 2025). Aunque algunos estudios utilizaron registros de plataformas virtuales como Moodle, se centraron más en el comportamiento digital que en las condiciones reales de acceso (Rebelo Marcolino et al., 2025).

Después de la pandemia surgieron nuevos factores como problemas emocionales, adaptación a la virtualidad y cambios en la motivación (Gonzalez et al., 2025). Sin embargo, pocos modelos actuales incluyen estas variables, lo que limita su validez (Duro et al., 2026). Por ello, se requieren modelos más integrales que combinen factores académicos, emocionales y tecnológicos, lo que permitirá una mejor comprensión de la deserción estudiantil (Tong & Li, 2025).

3. Necesidad de validar modelos en contextos institucionales específicos

El tercer vacío identificado se relaciona con la generalización excesiva de los modelos predictivos. Gran parte de los estudios revisados utilizan bases de datos nacionales, muestras agregadas o información de una sola institución, y luego extienden sus resultados a otros contextos sin realizar procesos rigurosos de validación externa

Esta práctica resulta problemática, ya que las causas de la deserción pueden variar de forma considerable entre universidades públicas y privadas, instituciones urbanas o rurales, carreras técnicas o profesionales, e incluso entre modalidades presencial, virtual o híbrida. Factores como las políticas de retención, el perfil socioeconómico del estudiantado, la calidad docente, los recursos institucionales y el entorno cultural influyen de manera distinta en cada organización educativa (Kocsis & Molnár, 2025).

En Ecuador, por ejemplo, los estudios disponibles muestran resultados diversos entre instituciones públicas, privadas y tecnológicas (Granda et al., 2024), lo que evidencia que no existen factores universales aplicables a todos los casos (Avilés-González et al., 2025). Situaciones similares también se observan en otros países latinoamericanos.

Diversos autores como Daza Santacoloma & Hoyos Osorio (2023) y Chacha et al. (23), coinciden en que los modelos predictivos deben calibrarse y validarse con datos propios de cada institución para garantizar precisión y utilidad práctica. Un modelo construido con información externa puede presentar bajo rendimiento cuando se aplica en contextos diferentes (Daza Santacoloma & Hoyos Osorio, 2023).

En este sentido, futuras investigaciones deberían priorizar el diseño de modelos adaptados a las realidades específicas de cada institución, complementados con validaciones cruzadas entre universidades con características semejantes. Esto permitiría desarrollar herramientas más confiables para la detección temprana del riesgo de deserción.

5. Propuesta de un modelo predictivo basado en regresión lineal múltiple (deserción universitaria)

Como propuesta práctica, se desarrolló un modelo predictivo utilizando la herramienta Anaconda, empleando un dataset público que se aproxima a las variables identificadas como relevantes en la literatura revisada. Entre ellas se consideraron factores académicos, personales y socioeconómicos vinculados a la deserción estudiantil. El objetivo fue evaluar la eficiencia del modelo mediante métricas estadísticas obtenidas durante el análisis. Este trabajo se complementa con el archivo Jupyter Notebook Propuesta_de_modelamiento, donde se detalla el desarrollo completo del modelo propuesto.

Descripción del dataset

Para el desarrollo de este estudio se utilizó un dataset público con información de estudiantes universitarios, el cual incluye variables relacionadas con el rendimiento académico, características personales y aspectos económicos, lo que lo hace adecuado para analizar el riesgo de deserción. La selección de este conjunto de datos se realizó a partir de la revisión de estudios previos, donde se identificaron las variables más relevantes para la predicción de la deserción

estudiantil, por lo que se eligió una base de datos con una estructura adecuada y suficiente información para construir el modelo. Entre las variables principales se consideran las calificaciones de ingreso, la edad, el estado de pago de matrícula, el número de materias aprobadas y la condición de beca, entre otras, las cuales fueron organizadas en categorías académicas, personales y socioeconómicas para facilitar su análisis.

Categoría	Variable en Dataset	Traducción / Significado
Personal	Age at enrollment	Edad al momento de ingreso
	Gender	Género del estudiante
	Marital status	Estado civil
Académica	Admission grade	Nota de admisión
	Previous qualification (grade)	Calificación académica previa
	Curricular units 1st sem (approved)	Materias aprobadas 1er semestre
	Curricular units 1st sem (grade)	Promedio 1er semestre
	Curricular units 2nd sem (approved)	Materias aprobadas 2do semestre
	Curricular units 2nd sem (grade)	Promedio 2do semestre
Socioeconómica	Debtor	Mantiene deuda pendiente
	Tuition fees up to date	Pagos al día
	Scholarship holder	Posee beca estudiantil
	Father's occupation	Ocupación del padre
	Mother's occupation	Ocupación de la madre

Figura

2. Tabla de variables utilizadas en el modelo clasificadas por tipo.

2. Fases del desarrollo del modelo

Para la construcción del modelo predictivo, se siguió un proceso organizado en tres fases:

■ Preprocesamiento de datos

En esta primera fase se trabajó con el dataset con el fin de prepararlo antes de aplicar el modelo. Primero, se cargaron los datos y se revisó su estructura para identificar la cantidad de registros y variables, lo que permitió comprender mejor la información disponible. Posteriormente, se realizó la limpieza de datos, corrigiendo valores mal formateados, además de revisar la presencia de datos faltantes y aplicar técnicas para completarlos, evitando posibles errores en el modelo. Luego, se transformó la variable objetivo asignándole valores numéricos que representaran el nivel de riesgo de deserción, ya que la regresión lineal requiere datos numéricos. Finalmente, se seleccionaron las variables más relevantes para el análisis, priorizando aquellas relacionadas con el rendimiento académico, la situación económica y ciertas características personales del estudiante.

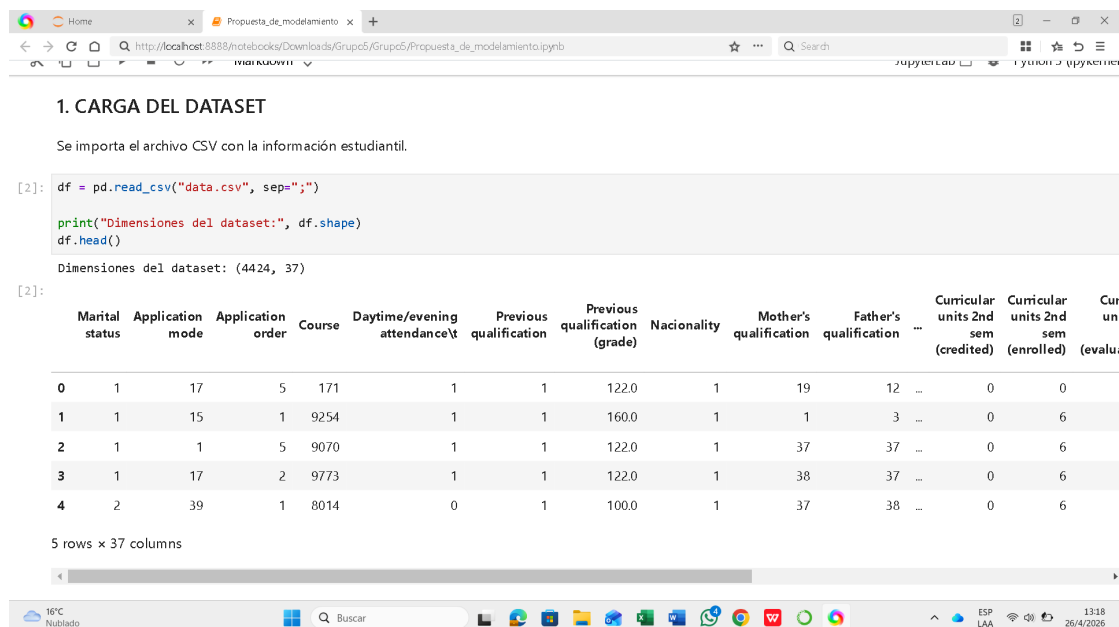


Figura 3. Visualización inicial del dataset.

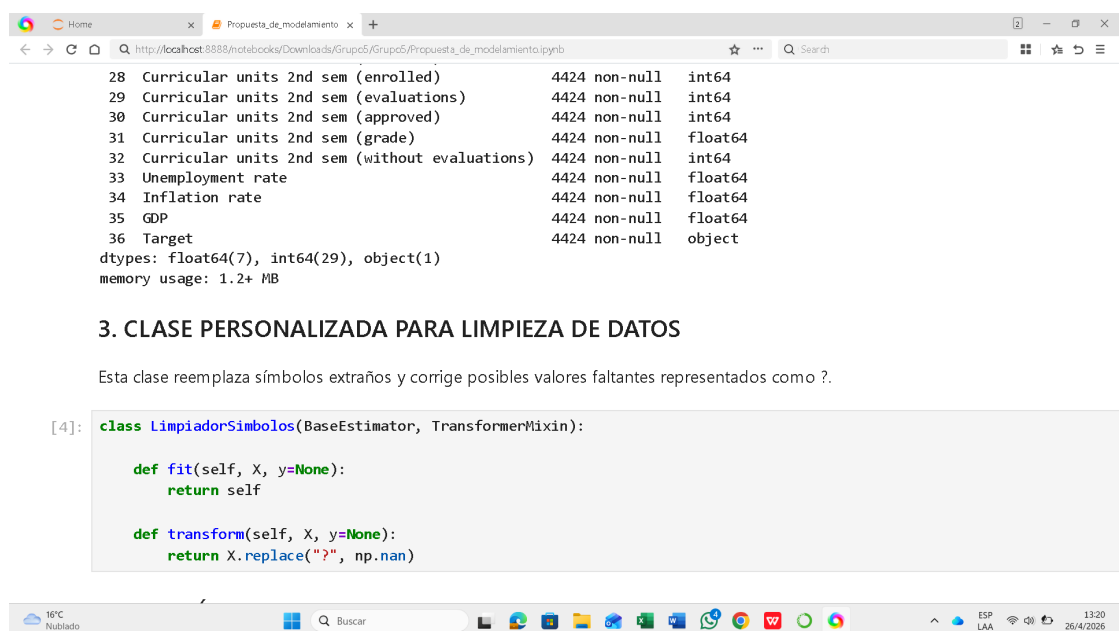


Figura 4. Proceso de limpieza de datos.

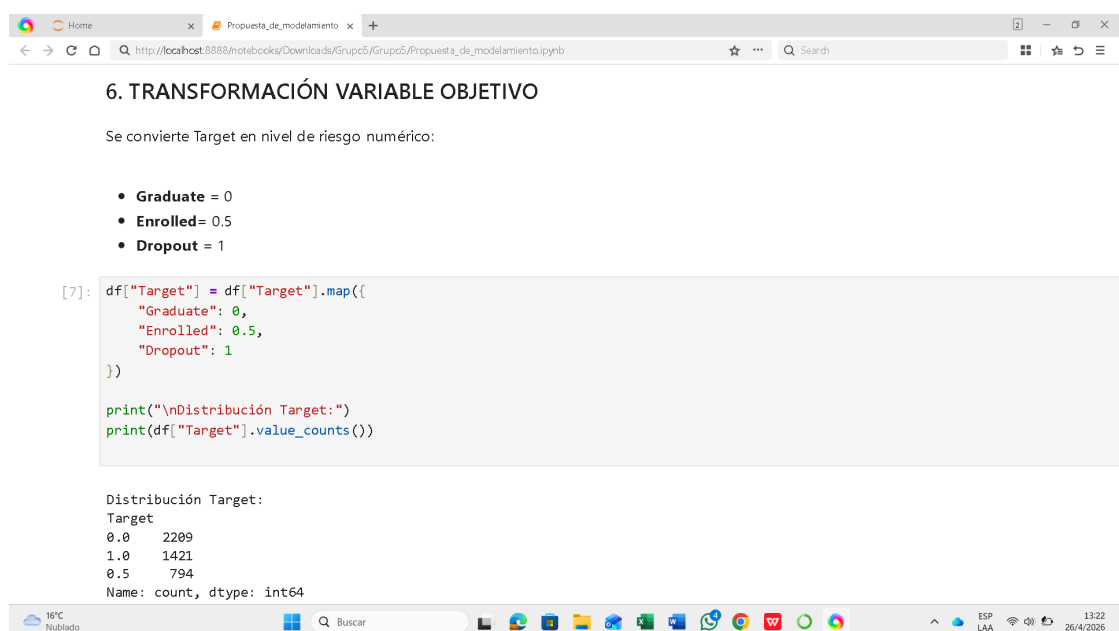
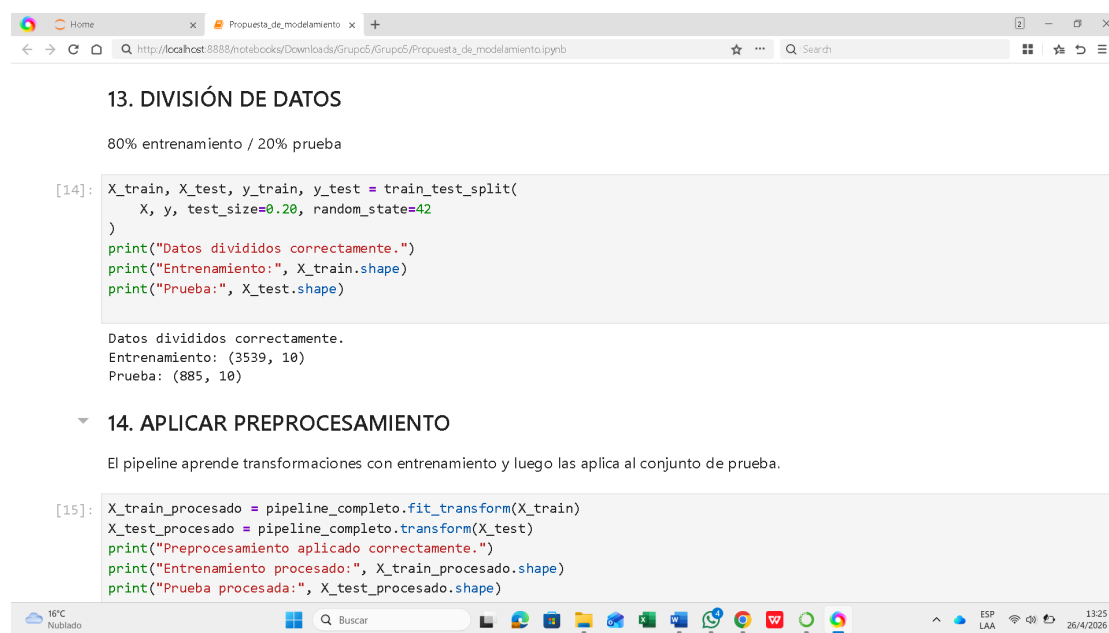


Figura 5. Transformación de la variable objetivo.

■ Construcción del modelado

En esta fase se construyó el modelo encargado de realizar la predicción. Primero, se dividieron los datos en dos conjuntos, uno de entrenamiento y otro de prueba, con el fin de evaluar el desempeño del modelo en datos que no ha visto previamente, lo que permite verificar su capacidad de generalización. Posteriormente, se aplicó un pipeline de preprocesamiento para organizar y preparar la información de manera automática, asegurando que los datos estuvieran en condiciones adecuadas para el modelo. Finalmente, se entrenó el modelo de regresión lineal múltiple utilizando las variables seleccionadas, lo que permitió analizar el efecto conjunto de diferentes factores sobre el riesgo de deserción estudiantil.



The screenshot shows a Jupyter Notebook interface with two cells. The first cell, titled '13. DIVISIÓN DE DATOS', contains code to split the data into training and testing sets using `train_test_split`. The output shows the data was split correctly with training set dimensions (3539, 10) and testing set dimensions (885, 10). The second cell, titled '14. APLICAR PREPROCESAMIENTO', contains code to apply a preprocessing pipeline to both training and testing sets. The output shows the preprocessing was applied correctly, resulting in processed training set dimensions (3539, 10) and processed testing set dimensions (885, 10).

```
[14]: X_train, X_test, y_train, y_test = train_test_split(
      X, y, test_size=0.20, random_state=42
    )
      print("Datos divididos correctamente.")
      print("Entrenamiento:", X_train.shape)
      print("Prueba:", X_test.shape)

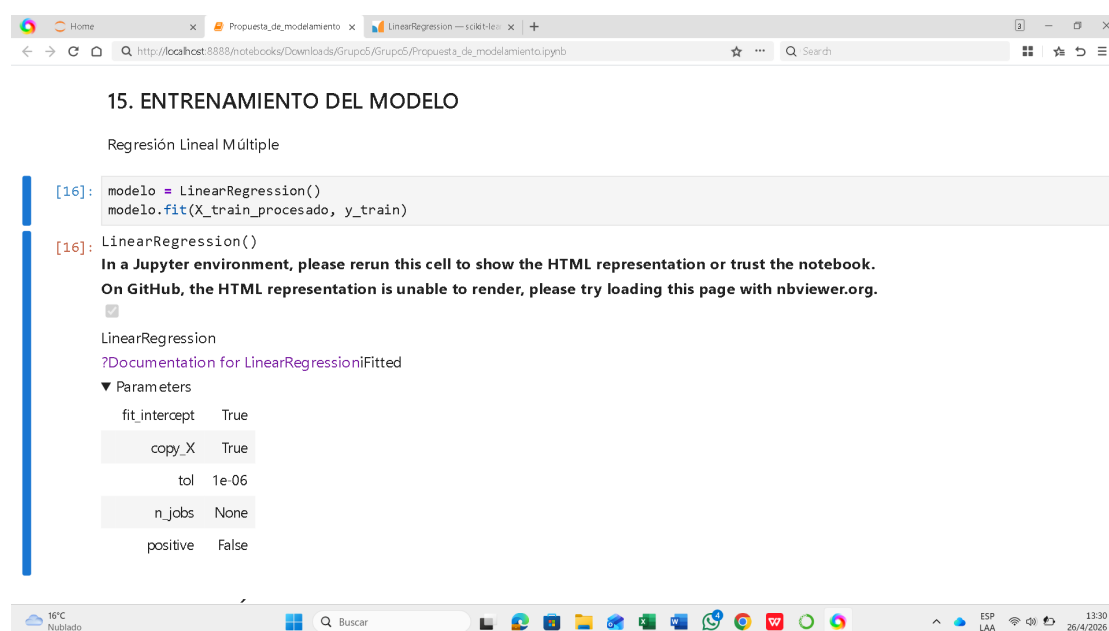
Datos divididos correctamente.
Entrenamiento: (3539, 10)
Prueba: (885, 10)
```

```
▼ 14. APLICAR PREPROCESAMIENTO

El pipeline aprende transformaciones con entrenamiento y luego las aplica al conjunto de prueba.

[15]: X_train_procesado = pipeline_completo.fit_transform(X_train)
      X_test_procesado = pipeline_completo.transform(X_test)
      print("Preprocesamiento aplicado correctamente.")
      print("Entrenamiento procesado:", X_train_procesado.shape)
      print("Prueba procesada:", X_test_procesado.shape)

Preprocesamiento aplicado correctamente.
Entrenamiento procesado: (3539, 10)
Prueba procesada: (885, 10)
```

Figura 6. División de los datos en entrenamiento y prueba.


The screenshot shows a Jupyter Notebook interface with two cells. The first cell, titled '15. ENTRENAMIENTO DEL MODELO', contains code to create a `LinearRegression` model and fit it to the training data. The output shows the model's parameters: `fit_intercept: True`, `copy_X: True`, `tol: 1e-06`, `n_jobs: None`, and `positive: False`.

```
[16]: modelo = LinearRegression()
      modelo.fit(X_train_procesado, y_train)
```

```
[16]: LinearRegression()
      In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
      On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.
      LinearRegression
      ?Documentation for LinearRegression
      ▼ Parameters
      fit_intercept True
      copy_X True
      tol 1e-06
      n_jobs None
      positive False
```

Figura 7. Entrenamiento del modelo de regresión lineal.

■ Evaluación del modelo

En esta última fase se evaluó el desempeño del modelo utilizando métricas como el R^2 y el RMSE. El R^2 permite conocer qué tanto el modelo explica el comportamiento del problema, mientras que el RMSE indica el nivel de error en las predicciones. En este caso, se obtuvo un R^2 de 0.4715, lo que representa una capacidad explicativa del 47.15 %, evidenciando un ajuste moderado del modelo. Además, se compararon los resultados entre los datos de entrenamiento y prueba para verificar si existía sobreajuste o un correcto comportamiento general. También se analizaron los coeficientes del modelo con el objetivo de identificar las variables con mayor influencia en el riesgo de deserción. Finalmente, se generó una gráfica comparando los valores reales con los predichos, lo que permitió visualizar el desempeño del modelo de forma más clara.

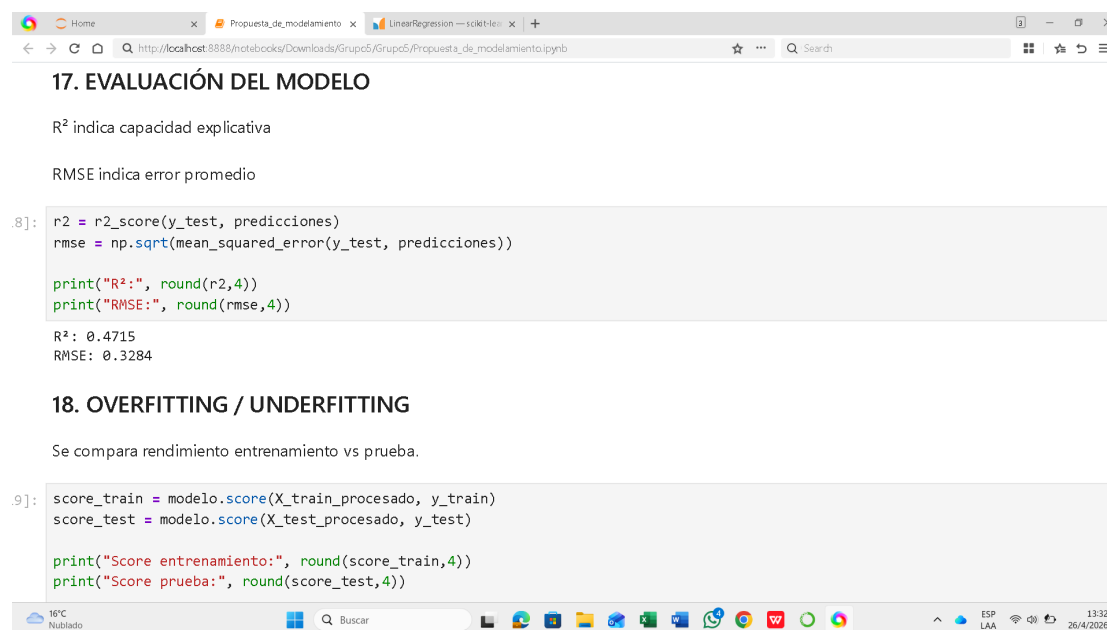


Figura 8. Resultados de las métricas de evaluación.

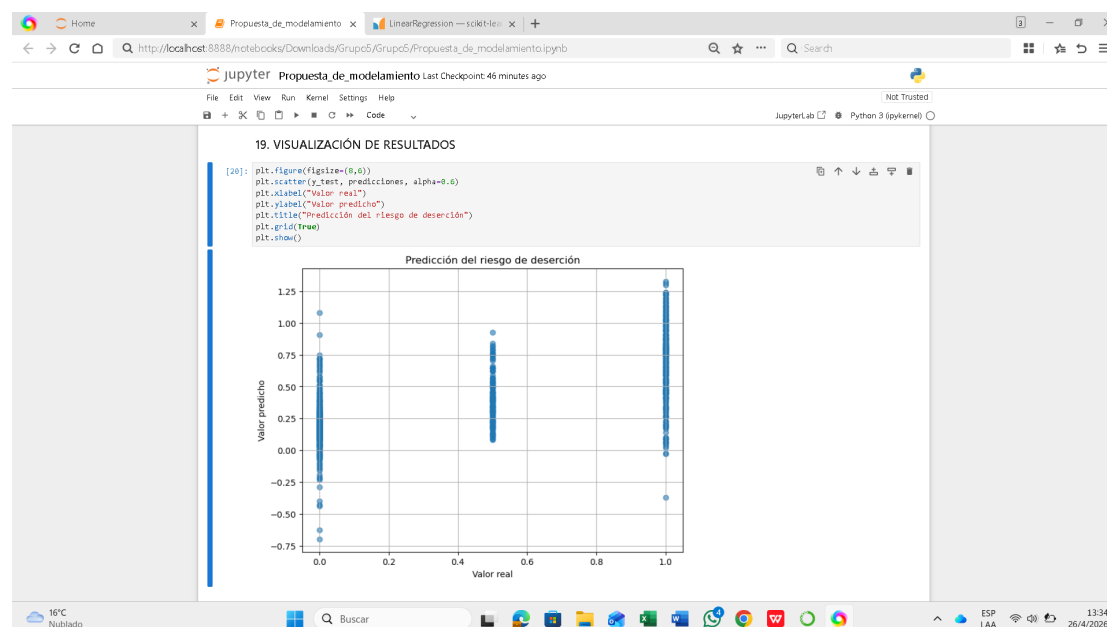


Figura 9. Comparación entre valores reales y predichos.

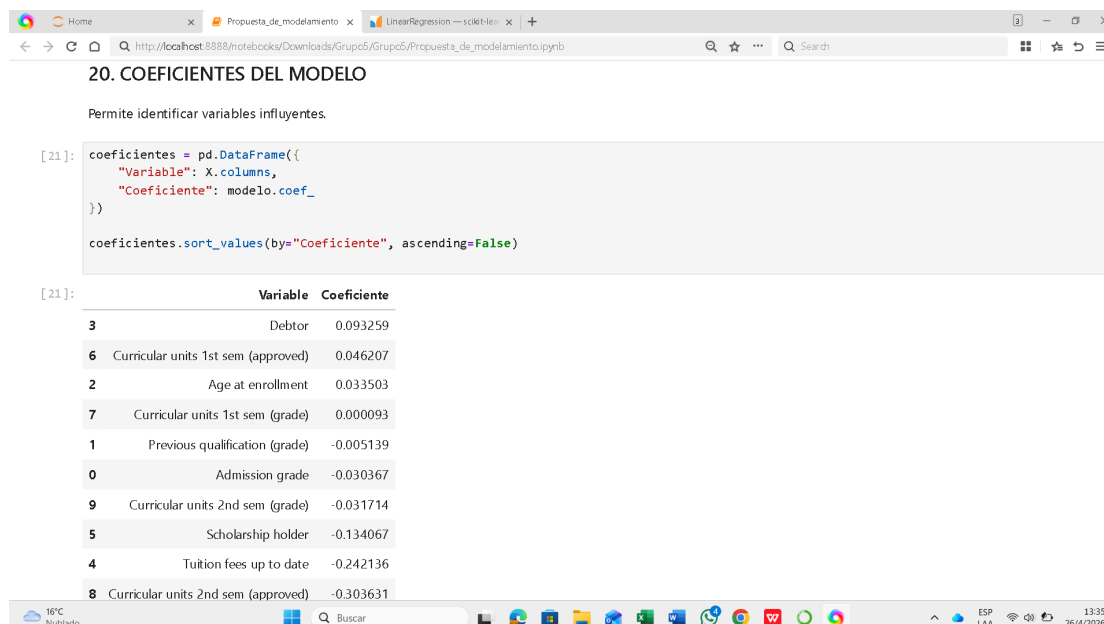


Figura 10. Importancia de las variables en el modelo.

Aplicaciones prácticas o futuras

La presente investigación bibliográfica, basada en una revisión sistemática con la metodología PRISMA, evidencia que los modelos de regresión lineal múltiple representan una herramienta útil en la predicción de la deserción estudiantil en la educación superior (Forti et al., 2025), ya que integran de manera simultánea diversos factores académicos, socioeconómicos y demográficos, orientados a la detección temprana de estudiantes con mayor riesgo de abandono, lo que facilita la planificación de intervenciones preventivas y focalizadas (Rabelo & Zárata, 2025).

Asimismo, la interpretabilidad de los modelos de regresión, en comparación con enfoques predictivos más complejos, facilita su aplicación en contextos institucionales reales (Duro et al., 2026), lo que resulta fundamental para que las autoridades académicas y los gestores educativos puedan diseñar políticas de retención basadas en evidencia empírica clara y comprensible (Romero & Liao, 2025).

En cuanto a las futuras líneas de investigación, se plantea como una proyección relevante la integración de modelos de regresión con técnicas de explicabilidad como SHAP y LIME, con el fin de reforzar la transparencia, la rendición de cuentas y el uso ético de los modelos predictivos en la educación superior (Duro et al., 2026). Estudios recientes señalan la importancia de evolucionar desde modelos enfocados únicamente en la precisión hacia sistemas predictivos que sean interpretables y útiles para la toma de decisiones (Quimiz-Moreira et al., 2025).

Adicionalmente, futuras investigaciones podrían profundizar en la incorporación de datos provenientes de plataformas virtuales de aprendizaje y sistemas de gestión académica, lo que permitiría enriquecer los modelos de regresión tradicionales con información sobre la interacción y el comportamiento estudiantil en entornos digitales, especialmente en modalidades de educación en línea y educación mixta (Rebelo Marcolino et al., 2025).

Finalmente, se propone el desarrollo de estudios longitudinales que sigan a los estudiantes a lo largo del tiempo, con el fin de analizar de forma más completa la interacción entre factores individuales, familiares e institucionales. Esto permitiría fortalecer tanto la capacidad predictiva como la explicativa de los modelos de regresión aplicados en la educación superior (Paksi et al., 2023).

4. CONCLUSIONES

Se confirma que la regresión lineal múltiple es una herramienta importante para comprender y predecir la deserción estudiantil, no solo por su capacidad para estimar el riesgo, sino también porque permite identificar con claridad los factores que más influyen en el abandono. Esto resulta especialmente útil para los investigadores del área educativa, ya que facilita un análisis más preciso del fenómeno, y para las instituciones, porque posibilita la aplicación de estos resultados en acciones concretas, como sistemas de alerta temprana o programas de apoyo dirigidos a estudiantes en riesgo.

Al mismo tiempo, los resultados refuerzan que la deserción es un fenómeno multifactorial, donde el rendimiento académico y las condiciones económicas tienen un peso importante, aunque no actúan de forma aislada. Este hallazgo amplía lo encontrado en estudios previos al integrar distintas variables en un mismo análisis, mostrando una visión más completa del problema que puede ser comprendida no solo por especialistas, sino también por docentes y autoridades educativas.

Otro aspecto relevante es que, aunque existen modelos más avanzados, la regresión lineal múltiple sigue siendo una opción práctica y accesible, lo que permite que investigadores de otros campos o instituciones con menos recursos puedan utilizarla sin dificultad. Esto facilita su aplicación en distintos contextos y acerca el uso del análisis de datos al público en general, especialmente en el ámbito educativo donde se requieren soluciones claras y aplicables.

Por otra parte, la evaluación del modelo muestra un comportamiento estable entre los datos de entrenamiento y prueba, lo que sugiere la ausencia de problemas significativos de sobreajuste. No obstante, el nivel de precisión alcanzado indica que la deserción estudiantil es un fenómeno complejo que no puede ser explicado en su totalidad mediante las variables consideradas. En este sentido, el modelo desarrollado constituye una aproximación inicial útil para la identificación temprana de estudiantes en riesgo, aunque se recomienda la incorporación de nuevas variables y el uso de técnicas más avanzadas en futuros estudios para mejorar su capacidad predictiva.

En conclusión, la aplicación del modelo de regresión lineal múltiple permitió estimar el riesgo de deserción estudiantil a partir de variables académicas, personales y socioeconómicas. Los resultados obtenidos evidencian que, si bien el modelo presenta una capacidad explicativa moderada, logra identificar relaciones relevantes entre factores como el rendimiento académico, el cumplimiento en el pago de matrícula y la condición de beca con el abandono estudiantil. Esto confirma la importancia de considerar múltiples dimensiones en el análisis de la deserción, en concordancia con lo reportado en la literatura.

FINANCIACIÓN

La presente investigación se desarrolló como parte de un trabajo académico correspondiente al sexto semestre de la carrera de Tecnologías de la Información de la Facultad de Informática y Electrónica de la Escuela Superior Politécnica de Chimborazo. La investigación fue realizada bajo la guía del Ingeniero correspondiente de la materia Machine Learning, con fines estrictamente académicos, y no cuenta con financiamiento externo proveniente de entidades públicas, privadas, organizaciones gubernamentales, becas ni proyectos de investigación financiados, por lo tanto, todos los recursos utilizados para el desarrollo de la investigación fueron de carácter académico e institucional.

CONFLICTO DE INTERESES

Durante el desarrollo de la presente investigación se declara que no se ha presentado ningún tipo de conflicto de interés que pudieran influir directa o indirectamente con el desarrollo de esta, puesto que el trabajo fue realizado con fines estrictamente académicos, como parte de formación universitaria.

CONTRIBUCIÓN DE AUTORÍA

Propuesta de distribución CRediT pendiente de confirmación expresa por los seis autores.

Tabla

Contribución de autoría según taxonomía CRediT

Participación	A1	A2	A3	A4	A5	A6
Conceptualización	X	X	X	X	X	X
Análisis formal	X	X	X	X	–	–
Investigación	X	X	X	X	X	X
Metodología	X	X	X	X	–	–
Recursos	–	X	X	–	X	X
Administración del proyecto	X	X	–	–	–	–
Redacción – borrador original	X	X	X	X	X	X
Redacción – revisión y edición	X	X	X	X	X	X
Discusión de resultados	X	X	X	X	X	X
Revisión y aprobación de la versión final	X	X	X	X	X	X

Nota: A1–A6 corresponden al orden de aparición de los autores. Las asignaciones son una propuesta editorial y deben validarse antes de publicar.

REFERENCIAS

- Aguilar-Reyes, J. E., Mejía-Peñañiel, E. F., Morocho-Barrionuevo, T. P., Velasco Castelo, G.-M., Aguilar-Reyes, J. E., Mejía-Peñañiel, E. F., Morocho-Barrionuevo, T. P., & Velasco Castelo, G.-M. (2025). ESTUDIO DEL RENDIMIENTO ACADÉMICO MEDIANTE LA COMPARACIÓN DE MODELOS DE REGRESIÓN Y ÁRBOLES DE CLASIFICACIÓN. *Telos*, 27(1), 94-115. <https://doi.org/10.36390/telos271.08>
- Andrade-Girón, D., Sandivar-Rosas, J., Marín-Rodriguez, W., Susanibar-Ramirez, E., Toro-Dextre, E., Ausejo-Sanchez, J., Villarreal-Torres, H., & Angeles-Morales, J. (2023). Predicting Student Dropout based on Machine Learning and Deep Learning: A Systematic Review. *EAI Endorsed Transactions on Scalable Information Systems*, 10(5). <https://doi.org/10.4108/eetsis.3586>
- Avilés-González, J., Córdova-León, F., Duque-Espinoza, G., & Mosquera-Gutierrez, J. (2025). Key Determinants of University Student Dropout: Insights from a Private Institution in Ecuador. *Journal of Latinos and Education*, 0(0), 1-16. <https://doi.org/10.1080/15348431.2025.2566917>
- Barragán Moreno, S. P., & Guzmán Rincón, A. (2025). Digital divide as an explanatory variable for dropout in higher education. *International Journal of Educational Technology in Higher Education*, 22(1), 60. <https://doi.org/10.1186/s41239-025-00550-0>
- Bernal Troya, A. S. (2025). Análisis predictivo de la permanencia y deserción estudiantil: Implementación del modelo de Churn en el Instituto Superior Tecnológico Wissen (2021-2024) [Tesis de Maestría en Inteligencia de Negocios y Ciencia de Datos, Universidad de las Américas]. <https://dspace.udla.edu.ec/bitstream/33000/17599/1/UDLA-EC-TMINCD-2025-20.pdf>
- Bettahi, A., Belouadha, F.-Z., & Harroud, H. (2025). A Modular and Explainable Machine Learning Pipeline for Student Dropout Prediction in Higher Education. *Algorithms*, 18(10), 662. <https://doi.org/10.3390/a18100662>
- Bouihi, B., Bousselham, A., Aoula, E., Ennibras, F., & Deraoui, A. (2024). Prediction of Higher Education Student Dropout based on Regularized Regression Models. *Engineering, Technology & Applied Science Research*, 14(6), 17811-17815. <https://doi.org/10.48084/etasr.8644>

- Chacha, B. R. C., López, W. L. G., & Constante, M. B. P. (2023). Predictive Model of Student Dropout Based on Logistic Regression. *ESPOCH Congresses: The Ecuadorian Journal of S.T.E.A.M.*, 630-656. <https://doi.org/10.18502/epoch.v3i1.14477>
- Chui B., H., Pérez A., K., Yabar M., P., Chui B., H., Pérez A., K., & Yabar M., P. (2025). Factores que intervienen en la deserción estudiantil universitaria. *Revista de Investigaciones Veterinarias del Perú*, 36(3). <https://doi.org/10.15381/rivep.v36i3.30906>
- Contreras, D., González, L., Láscar, S., & López, V. (2022). Negative teacher–student and student–student relationships are associated with school dropout: Evidence from a large-scale longitudinal study in Chile. *International Journal of Educational Development*, 91, 102576. <https://doi.org/10.1016/j.ijedudev.2022.102576>
- Dang, T. K., & Nguyen, H. H. X. (2022). A Hybrid Approach Using Decision Tree and Multiple Linear Regression for Predicting Students' Performance Based on Learning Progress and Behavior. *SN Computer Science*, 3(5), 393. <https://doi.org/10.1007/s42979-022-01251-5>
- Daza Santacoloma, G., & Hoyos Osorio, J. K. (2023). Predictive Model to Identify College Students with High Dropout Rates. *Revista Electrónica de Investigación Educativa*, 25. <https://doi.org/10.24320/redie.2023.25.e13.5398>
- Duro, B., Gomes, A., Correia, F. B., Borges, A. R., & Bernardino, J. (2026). Machine Learning and Deep Learning for Dropout Prediction in Higher Education: A Review. *Computers*, 15(3), 164. <https://doi.org/10.3390/computers15030164>
- Forti, M., Bilancia, M., Cafarelli, B., del Gobbo, E., & Nigri, A. (2025). Predicting dropout from higher education with multinomial finite mixture models: Empirical evidence from Italy. *Quality & Quantity*. <https://doi.org/10.1007/s11135-025-02135-5>
- Garrido Silva, C. A., Pajuelo Díaz, J., Garrido Silva, C. A., & Pajuelo Díaz, J. (2023). Deserción en estudiantes de educación superior: Un estudio de caso. *Universidad, Ciencia y Tecnología*, 27(119), 18-28. <https://doi.org/10.47460/uct.v27i119.703>
- Gonzalez, B., Mendes, T. P., Pinto, R., Correia, S. V., Albuquerque, S., & Paulino, P. (2025). Predictors of higher education dropout intention in the post-pandemic era: The mediating role of academic exhaustion. *PLOS ONE*, 20(7), e0327643. <https://doi.org/10.1371/journal.pone.0327643>
- Granda, L. E. Á., Yautibug, F. C., & Copa, R. A. (2024). Deserción en la Educación Superior en Ecuador, Causas y Consecuencias. *Ciencia Latina Revista Científica Multidisciplinar*, 8(3), 11475-11490. https://doi.org/10.37811/cl_rcm.v8i3.12472
- Huynh-Cam, T.-T., Chen, L.-S., & Lu, T.-C. (2025). Early Prediction Models and Crucial Factor Extraction for First-Year Undergraduate Student Dropouts. *Journal of Applied Research in Higher Education*, 17(2), 624-639. <https://doi.org/10.1108/JARHE-10-2023-0461>
- Jiménez-Gutiérrez, A. L., Mota-Hernández, C. I., Mezura-Montes, E., & Alvarado-Corona, R. (2024). Application of the performance of machine learning techniques as support in the prediction of school dropout. *Scientific Reports*, 14(1), 3957. <https://doi.org/10.1038/s41598-024-53576-1>
- Kocsis, Á., & Molnár, G. (2025). Factors influencing academic performance and dropout rates in higher education. *Oxford Review of Education*, 51(3), 414-432. <https://doi.org/10.1080/03054985.2024.2316616>
- Lima, M. N. C. A., & Fagundes, R. A. de A. (2020). Educational DataMining: A Study of the Factors That Cause School Dropout in Higher Education Institutions in Brazil. *RENOTE*, 18(1). <https://doi.org/10.22456/1679-1916.105950>
- Marca Maquera, H., Coyla Zela, M., & Paredes, R. P. (2025). Factors at the household and university level that influence student dropout at UNAM. *Frontiers in Education*, 10, 1598687. <https://doi.org/10.3389/feduc.2025.1598687>

- Márquez, F., & Sánchez-Pozo, N. (2023). Modelo de predicción de deserción: Un estudio de caso de estudiantes de la Universidad Politécnica Estatal del Carchi. *Prometeo Conocimiento Científico*, 3, e34. <https://doi.org/10.55204/pcc.v3i2.e34>
- Núñez-Naranjo, A. F. (2024a). Analysis of the determinant factors in university dropout: A case study of Ecuador. *Frontiers in Education*, 9, 1444534. <https://doi.org/10.3389/feduc.2024.1444534>
- Núñez-Naranjo, A. F. (2024b). Analysis of the determinant factors in university dropout: A case study of Ecuador. *Frontiers in Education*, 9. <https://doi.org/10.3389/feduc.2024.1444534>
- Orozco-Rodríguez, C., Viegas, C., Costa, A. R., Lima, N., & Alves, G. R. (2025). Dropout Rate Model Analysis at an Engineering School. *Education Sciences*, 15(3), 287. <https://doi.org/10.3390/educsci15030287>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021a). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *PLoS Medicine*, 18(3), e1003583. <https://doi.org/10.1371/journal.pmed.1003583>
- Paksi, B., Széll, K., & Fehérvári, A. (2023). Empirical Testing of a Multidimensional Model of School Dropout Risk. *Social Sciences*, 12(2), 50. <https://doi.org/10.3390/socsci12020050>
- Pavón, M. A. L., & Paez, S. A. C. (2025). Modelado del índice de deserción estudiantil en Carreras de Ingeniería en una Institución de Educación Superior.: Modeling of the student dropout rate in Engineering Degrees in a Higher Education Institution. *Revista Científica Multidisciplinar G-nerando*, 6(2), ág. 228-242. <https://doi.org/10.60100/rcmg.v6i2.733>
- Quimiz-Moreira, M., Delgadillo, R., Parraga-Alava, J., Maculan, N., & Mauricio, D. (2025). Factors, Prediction, Explainability, and Simulating University Dropout Through Machine Learning: A Systematic Review, 2012–2024. *Computation*, 13(8), 198. <https://doi.org/10.3390/computation13080198>
- Rabelo, A. M., & Zárate, L. E. (2025). A model for predicting dropout of higher education students. *Data Science and Management*, 8(1), 72-85. <https://doi.org/10.1016/j.dsm.2024.07.001>
- Rahmani, A. M., Groot, W., & Rahmani, H. (2024). Dropout in online higher education: A systematic literature review. *International Journal of Educational Technology in Higher Education*, 21(1), 19. <https://doi.org/10.1186/s41239-024-00450-9>
- Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting Student Dropout and Academic Success. *Data*, 7(11), 146. <https://doi.org/10.3390/data7110146>
- Rebelo Marcolino, M., Reis Porto, T., Thompsons Primo, T., Targino, R., Ramos, V., Marques Queiroga, E., Munoz, R., & Cechinel, C. (2025). Student dropout prediction through machine learning optimization: Insights from moodle log data. *Scientific Reports*, 15(1), 9840. <https://doi.org/10.1038/s41598-025-93918-1>
- Romero, S., & Liao, X. (2025). Statistical and machine learning models for predicting university dropout and scholarship impact. *PLOS ONE*, 20(6), e0325047. <https://doi.org/10.1371/journal.pone.0325047>
- Segura, M., Mello, J., & Hernández, A. (2022). Machine Learning Prediction of University Student Dropout: Does Preference Play a Key Role? *Mathematics*, 10(18), 3359. <https://doi.org/10.3390/math10183359>
- Selim, K. S., & Rezk, S. S. (2023). On predicting school dropouts in Egypt: A machine learning approach. *Education and Information Technologies*, 28(7), 9235-9266. <https://doi.org/10.1007/s10639-022-11571-x>

- Seo, E.-Y., Yang, J., Lee, J.-E., & So, G. (2024). Predictive modelling of student dropout risk: Practical insights from a South Korean distance university. *Heliyon*, 10(11). <https://doi.org/10.1016/j.heliyon.2024.e30960>
- Song, Z., Sung, S.-H., Park, D.-M., & Park, B.-K. (2023). All-Year Dropout Prediction Modeling and Analysis for University Students. *Applied Sciences*, 13(2), 1143. <https://doi.org/10.3390/app13021143>
- Tong, T., & Li, Z. (2025). Predicting learning achievement using ensemble learning with result explanation. *PLOS ONE*, 20(1), e0312124. <https://doi.org/10.1371/journal.pone.0312124>
- 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>