Instituto Internacional de Investigación y Desarrollo Tecnológico Educativo INDTEC, C.A.
DOI: https://doi.org/10.29394/Scientific.issn.2542-2987.2026.11.40.6.93-111
OAI-PMH: http://www.indteca.com/ojs/index.php/Revista_Scientific/oai

Modelos Predictivos en Python para la Deserción Universitaria: Revisión Sistemática

Jesús Eduardo Zambrano Loor[1]; Luis Daniel Muñoz Mendoza[1]; Jodamia Uridisnalda Murillo Rosado[1]

[1]Universidad San Gregorio de Portoviejo (USGP), Manabí, Ecuador
jezambrano@sangregorio.edu.ec | https://orcid.org/0009-0004-9228-0237
ldmunoz@sangregorio.edu.ec | https://orcid.org/0009-0002-2130-7311
jumurillo@sangregorio.edu.ec | https://orcid.org/0000-0003-4558-5009
Correspondencia: jezambrano@sangregorio.edu.ec

Resumen: La deserción universitaria constituye un fenómeno multidimensional cuya predicción temprana mediante aprendizaje automático se perfila como una estrategia prometedora para las políticas de retención. El objetivo fue sistematizar la evidencia científica sobre modelos predictivos de aprendizaje automático, implementables en Python, para anticipar la deserción universitaria. Se empleó un enfoque cualitativo, método analítico-sintético y diseño de revisión sistemática conforme a PRISMA 2020. La búsqueda se realizó en Google Académico y SciELO, sin acceso a bases de suscripción, sobre publicaciones en español e inglés de 2014 a 2025. De 80 registros identificados, tras eliminar duplicados, cribar por título y resumen y evaluar a texto completo, se incluyeron 16 estudios nuevos que, sumados a 34 referencias de base, conformaron un corpus de 50 referencias temáticas. Los hallazgos muestran que los algoritmos de ensamble, como Random Forest, XGBoost y CatBoost, y las redes neuronales alcanzan una exactitud de entre 77,3% y 96,87%, o un área bajo la curva ROC (AUC) de hasta 0,957, según el contexto institucional, aunque su ventaja sobre la regresión logística no fue uniforme. La calidad del reporte fue alta (4,9 sobre 5), pero la evidencia muestra heterogeneidad metodológica, seis estudios comparten un mismo conjunto de datos y la producción latinoamericana se concentra en Perú y en tesis, sin estudios ecuatorianos. Se concluye que estos modelos son herramientas técnicamente prometedoras para anticipar la deserción universitaria, cuya eficacia depende de articularse con políticas institucionales éticas, transparentes y sostenidas en el tiempo.

Palabras clave: deserción universitaria; modelos predictivos; python; aprendizaje automático; revisión sistemática.
Código de clasificación internacional UNESCO: 5802.04 - Niveles y temas de educación.
Clasificación OCDE-FOS: 1.2 - Ciencias de la computación e información.

Predictive Models in Python for University Dropout: A Systematic Review

Abstract: University dropout constitutes a multidimensional phenomenon whose early prediction through machine learning is emerging as a promising strategy for retention policies. The objective was to systematize the scientific evidence on machine learning predictive models, implementable in Python, to anticipate university dropout. We used a qualitative approach, an analytical-synthetic method, and a systematic review design in accordance with PRISMA 2020. We conducted the search in Google Scholar and SciELO, without access to subscription databases, on publications in Spanish and English from 2014 to 2025. Of 80 records identified, after removing duplicates, screening by title and abstract, and assessing full texts, we included 16 new studies that, added to 34 baseline references, made up a corpus of 50 thematic references. The findings showed that ensemble algorithms, such as Random Forest, XGBoost, and CatBoost, and neural networks achieved accuracy ranging from 77.3% to 96.87%, or an area under the ROC curve (AUC) of up to 0.957, depending on the institutional context, although their advantage over logistic regression was not uniform. The reporting quality was high (4.9 out of 5), but the evidence showed methodological heterogeneity, six studies shared the same dataset, and Latin American research output was concentrated in Peru and in theses, with no Ecuadorian studies. We conclude that these models are technically promising tools for anticipating university dropout, whose effectiveness depends on their being combined with ethical, transparent, and sustained institutional policies.

Keywords: university dropout; predictive models; python; machine learning; systematic review.
UNESCO International Classification Code: 5802.04 - Levels and subjects of education.
OECD-FOS Classification: 1.2 - Computer and information sciences.

Modelos Preditivos em Python para a Evasão Universitária: Revisão Sistemática

Resumo: A evasão universitária constitui um fenômeno multidimensional cuja predição precoce por meio do aprendizado de máquina se configura como uma estratégia promissora para as políticas de permanência. O objetivo foi sistematizar a evidência científica sobre modelos preditivos de aprendizado de máquina, implementáveis em Python, para antecipar a evasão universitária. Empregou-se uma abordagem qualitativa, método analítico-sintético e delineamento de revisão sistemática conforme a PRISMA 2020. A busca foi realizada no Google Acadêmico e na SciELO, sem acesso a bases de assinatura, sobre publicações em espanhol e inglês de 2014 a 2025. Dos 80 registros identificados, após a remoção de duplicatas, a triagem por título e resumo e a avaliação do texto completo, foram incluídos 16 novos estudos que, somados a 34 referências de base, constituíram um corpus de 50 referências temáticas. Os achados mostram que os algoritmos de ensemble, como Random Forest, XGBoost e CatBoost, e as redes neurais alcançam uma acurácia de entre 77,3% e 96,87%, ou uma área sob a curva ROC (AUC) de até 0,957, conforme o contexto institucional, embora sua vantagem sobre a regressão logística não tenha sido uniforme. A qualidade do relato foi alta (4,9 sobre 5), mas a evidência mostra heterogeneidade metodológica, seis estudos compartilham um mesmo conjunto de dados, e a produção latino-americana se concentra no Peru e em teses, sem estudos equatorianos. Conclui-se que esses modelos são ferramentas tecnicamente promissoras para antecipar a evasão universitária, cuja eficácia depende de sua articulação com políticas institucionais éticas, transparentes e sustentadas no tempo.

Palavras-chave: evasão universitária; modelos preditivos; python; aprendizado de máquina; revisão sistemática.
Código de Classificação Internacional da UNESCO: 5802.04 - Níveis e temas de educação.
Classificação OCDE-FOS: 1.2 - Ciências da computação e informação.

Cómo citar este artículo:
Zambrano, J. E., Muñoz, L. D., & Murillo, J. U. (2026). Modelos Predictivos en Python para la Deserción Universitaria: Revisión Sistemática. *Revista Scientific, 11*(40), 93–111. https://doi.org/10.29394/Scientific.issn.2542-2987.2026.11.40.6.93-111

Fecha de Recepción:	Fecha de Aceptación:	Fecha de Publicación:
17-09-2025	08-11-2025	05-05-2026

Revista Scientific - Artículo de Revisión Sistemática - Arbitrado - Volumen 11, Número 40 - Mayo-Julio 2026 - pp. 93-111
pp. BA2016000002 - e-ISSN: 2542-2987 - ISNI: 0000 0004 6045 0361

DOI: https://doi.org/10.29394/Scientific.issn.2542-2987.2026.11.40.6.93-111
OAI-PMH: http://www.indteca.com/ojs/index.php/Revista_Scientific/oai

1. Introducción

La estabilidad estudiantil en la educación superior constituye uno de los pilares más importantes para medir los indicadores de calidad de una institución y su compromiso con la sociedad. Sin embargo, en América Latina los indicadores de deserción universitaria generan una intensa preocupación, y entre las causas que influyen en este fenómeno se encuentran las limitaciones económicas, el escaso acompañamiento académico, los retrasos en el ingreso al sistema educativo y los factores personales del estudiante vinculados con su proyecto de vida (Munizaga et al., 2018; Ramírez & Carcausto-Calla, 2024). Este fenómeno no solo representa una pérdida de oportunidades de estudio para el estudiante, sino que también repercute en las universidades y en el Estado, en cuanto a su inversión en la formación superior (Ramírez & Carcausto-Calla, 2024).

En los últimos años, el crecimiento acelerado de los datos en entornos universitarios ha abierto nuevas oportunidades para detectar, comprender y enfrentar la predicción del abandono estudiantil. Entre ellas, los modelos basados en *machine learning* o aprendizaje automático se han transformado en instrumentos de gran potencial, ya que identifican patrones ocultos en el rendimiento académico y permiten anticipar el abandono estudiantil, superando las limitaciones de los métodos tradicionales de seguimiento (Sharma et al., 2023; Shrestha & Pokharel, 2020).

De acuerdo con esto, Python, por su activa comunidad y su naturaleza de código abierto, dispone de bibliotecas maduras para el aprendizaje automático, como Scikit-learn (Pedregosa et al., 2011), que facilitan la creación de modelos de predicción. Esto abre la puerta para que las instituciones que todavía utilizan sistemas generales de detección y prevención no queden rezagadas en un escenario cada vez más orientado hacia la personalización de las medidas y la predictibilidad.

En este artículo, mediante una revisión sistemática de la literatura científica desarrollada conforme a las directrices PRISMA 2020, se analizan los fundamentos conceptuales y técnicos del uso de los modelos predictivos implementables en Python para anticipar la deserción estudiantil en las universidades. Asimismo, se destacan las oportunidades que ofrecen estas herramientas y los desafíos metodológicos, éticos y operativos que implica su utilización en la educación superior. En este trabajo, los términos deserción y abandono universitario se emplean como sinónimos, y exactitud designa la proporción de clasificaciones correctas *(accuracy)* que reporta cada estudio.

1.1. La Deserción Universitaria como Fenómeno Complejo y Multidimensional

Dentro de este fenómeno, la deserción universitaria puede entenderse como un proceso complejo y multidimensional que incluye factores académicos, socioeconómicos, personales e institucionales; para su predicción se requieren enfoques integrales, apoyados en modelos de aprendizaje automático y en el análisis minucioso de los datos disponibles; en este sentido, Sihare (2024) señala que constituye un problema persistente tanto en países desarrollados como en desarrollo y que los enfoques de inteligencia artificial y aprendizaje automático resultan eficaces para favorecer la retención.

Nicoletti y de Oliveira (2020) reconocen que el concepto de deserción universitaria ha sido definido desde múltiples perspectivas y proponen un sistema computacional basado en aprendizaje automático para predecirla a partir de registros históricos de abandono. En concordancia con ello, y con lo señalado por Munizaga et al. (2018), quienes agrupan las variables asociadas al abandono en factores individuales, académicos, económicos, institucionales y culturales y advierten que su reducción demanda cambios institucionales y garantías de financiamiento, la deserción debe entenderse como el resultado de un proceso en el que interactúan diferentes dimensiones, lo que exige instrumentos analíticos que permitan diseñar estrategias institucionales más efectivas para su reducción.

Revista Scientific - Artículo de Revisión Sistemática - Arbitrado - Volumen 11, Número 40 - Mayo-Julio 2026 - pp. 93-111
pp. BA2016000002 - e-ISSN: 2542-2987 - ISNI: 0000 0004 6045 0361

DOI: https://doi.org/10.29394/Scientific.issn.2542-2987.2026.11.40.6.93-111
OAI-PMH: http://www.indteca.com/ojs/index.php/Revista_Scientific/oai

Alves et al. (2024) caracterizan la deserción universitaria como un fenómeno multifactorial y señalan que los modelos predictivos y las intervenciones personalizadas favorecen la retención, pues permiten identificar de manera temprana a los estudiantes en riesgo y orientar las estrategias institucionales para evitar el abandono.

Por otro lado, de acuerdo con el estudio de Ramírez y Carcausto-Calla (2024), la deserción se asocia con factores académicos, individuales, económicos e institucionales, lo que respalda el diseño de estrategias de retención basadas en el apoyo académico y psicológico a los estudiantes. La detección temprana del abandono mediante técnicas de aprendizaje automático y minería de datos se sustenta, en cambio, en la literatura sobre minería de datos educativos y modelos predictivos de aprendizaje automático.

En América Latina, en los últimos años, la deserción estudiantil se ha visto recrudecida por la problemática sanitaria y por factores sociopolíticos propios de la región (Ramírez & Carcausto-Calla, 2024). A juicio de los autores del presente artículo, esta situación exige respuestas coordinadas entre las instituciones de educación superior y las políticas públicas para reducir sus efectos sociales y económicos.

De manera complementaria, Patacsil (2020) propone un enfoque de análisis de supervivencia *(survival analysis)* aplicado a datos de matrícula estudiantil, combinado con modelos de ensamble, para la predicción temprana de la deserción; en su estudio, el ensamble por *bagging* sobre árboles J48 obtuvo la mayor exactitud con los datos de matrícula, y este enfoque permite a las universidades identificar tempranamente a los estudiantes en riesgo a partir de la evolución de su trayectoria académica y fortalecer así las políticas de permanencia.

La deserción universitaria, según Zając et al. (2024), se asocia con la salud mental de los estudiantes: a partir de datos administrativos australianos, los autores muestran que recibir tratamiento por problemas de salud mental se relaciona con una mayor probabilidad de abandono, incluso al controlar por factores individuales, familiares y del programa de estudio. Este hallazgo refuerza la necesidad de políticas institucionales de prevención y acompañamiento temprano.

1.2. Minería de Datos Educativos y Modelos Predictivos de Aprendizaje Automático

Una vez caracterizada la deserción universitaria como un fenómeno multidimensional, resulta pertinente revisar los aportes de la minería de datos educativos y de los modelos predictivos de aprendizaje automático que han permitido operacionalizar su estudio.

Según Shrestha y Pokharel (2020), la minería de datos educativos ha surgido como un campo interdisciplinario clave, ya que convierte los enormes volúmenes de datos generados por la educación en información útil para la toma de decisiones institucionales en las instituciones de educación superior; su aplicación se orienta, sobre todo, a predecir el éxito académico mediante técnicas de clasificación, agrupamiento y asociación.

Por su parte, Tao (2024) propone un modelo de minería de datos con selección adaptativa de características y XGBoost que mejora la predicción del rendimiento estudiantil frente a métodos tradicionales, lo que ilustra el aporte de la selección de variables a la precisión de los modelos y subraya la importancia de adoptar marcos metodológicos rigurosos para incrementar la confianza en los modelos predictivos.

En este mismo sentido, Ouyang et al. (2023) muestran que integrar un modelo de predicción de desempeño con retroalimentación de analítica del aprendizaje aumentó el compromiso y mejoró el desempeño colaborativo en un curso de ingeniería en línea, lo que sugiere que la predicción resulta más útil cuando orienta acciones de acompañamiento para cada estudiante.

De acuerdo con Xu et al. (2022), un modelo de red neuronal convolucional que incorpora la correlación

Revista Scientific - Artículo de Revisión Sistemática - Arbitrado - Volumen 11, Número 40 - Mayo-Julio 2026 - pp. 93-111
pp. BA2016000002 - e-ISSN: 2542-2987 - ISNI: 0000 0004 6045 0361

DOI: https://doi.org/10.29394/Scientific.issn.2542-2987.2026.11.40.6.93-111
OAI-PMH: http://www.indteca.com/ojs/index.php/Revista_Scientific/oai

de la conducta de aprendizaje en días consecutivos predice el abandono en cursos abiertos masivos en línea (MOOC) con mayor exactitud que los métodos previos. La transparencia, la interpretabilidad y los principios éticos constituyen, a juicio de los autores del presente artículo, condiciones adicionales para que estos modelos se conviertan en herramientas confiables de apoyo a la toma de decisiones.

De acuerdo con Urbina-Nájera et al. (2021), el uso de minería de datos educativos en un estudio que abarcó más de 10.000 registros de una universidad privada en México permitió identificar con alta exactitud a los alumnos en situación de riesgo, mediante árboles de decisión conocidos como C4.5 y REPTree. Los modelos presentados no se limitan a alcanzar altos niveles de exactitud en su predicción, sino que también otorgan reglas claras y precisas, lo que permite a las instituciones planear estrategias de prevención y retención más efectivas.

También existe una deserción masiva en los cursos abiertos en línea (MOOC), fenómeno que puede anticiparse con mayor precisión mediante técnicas de aprendizaje automático, como muestran Ye y Biswas (2014), quienes en su investigación incorporan información temporal de mayor granularidad a los modelos de predicción y demuestran que este tipo de variables aumenta el poder predictivo en las fases tempranas de un curso. Estos resultados destacan la importancia de una ingeniería centrada en la identificación de las características y comportamientos de los estudiantes, lo que permite implementar estrategias de intervención temprana orientadas a mejorar la retención en los entornos en línea.

Cuando se implementan modelos predictivos basados en minería de datos, estos se consideran una herramienta estratégica para las instituciones de educación superior, en la medida en que buscan reducir la deserción y mejorar la retención estudiantil. En este sentido, el estudio realizado por Cholis y Ulinnuha (2023), fundamentado en el análisis de 1.092 registros de estudiantes de la Universitas Islam Negeri Sunan Ampel Surabaya, en Indonesia, demostró que un modelo de votación por conjunto *(ensemble voting)* que combina el árbol de decisión C4.5, K-Nearest Neighbor y *backpropagation* supera a cada uno de estos algoritmos por separado, alcanzando un 98,18% de exactitud en la clasificación de estudiantes desertores; estos resultados resaltan la importancia de combinar distintos algoritmos de aprendizaje automático en este tipo de investigaciones.

La investigación realizada por Perkash et al. (2024) señala la importancia de incorporar datos provenientes de varias fuentes, tanto a nivel académico como demográfico, los cuales pueden extraerse mediante plataformas como Moodle; esta combinación de datos permite mejorar la capacidad predictiva. Entre los factores más destacados se encuentran la cantidad de créditos aprobados, las asignaturas reprobadas y el nivel de actividad del estudiante dentro de la plataforma virtual. Según los autores, el peso de estas variables no es estático, sino que varía con el tiempo, lo que indica la necesidad de utilizar herramientas que incorporen variables temporales.

En esta dirección, Yükseltürk et al. (2014) aplicaron los algoritmos k-NN, árboles de decisión, redes neuronales y Naive Bayes a variables demográficas y de autoeficacia y preparación para el aprendizaje en línea de los estudiantes de un programa en línea, y el algoritmo k-NN alcanzó la mayor sensibilidad (87%) para identificar a los desertores, lo que muestra que la selección de las variables y del algoritmo incide en la capacidad de detección del abandono.

Por otro lado, la revisión sistemática de Murnawan et al. (2024) identifica que los datos demográficos, las calificaciones previas, los factores sociales y la actividad en línea se cuentan entre los factores más empleados para predecir el rendimiento estudiantil, lo que respalda la combinación de variables académicas, socioeconómicas y demográficas en los marcos metodológicos de predicción.

Asimismo, Shrestha y Pokharel (2021) aplican algoritmos de clasificación (KNN, Naive Bayes, SVM,

Revista Scientific - Artículo de Revisión Sistemática - Arbitrado - Volumen 11, Número 40 - Mayo-Julio 2026 - pp. 93-111
pp. BA2016000002 - e-ISSN: 2542-2987 - ISNI: 0000 0004 6045 0361

DOI: https://doi.org/10.29394/Scientific.issn.2542-2987.2026.11.40.6.93-111
OAI-PMH: http://www.indteca.com/ojs/index.php/Revista_Scientific/oai

Random Forest y CART) a datos de Moodle para identificar las variables más influyentes y facilitar la detección temprana del rendimiento estudiantil, y las máquinas de vectores de soporte (SVM) alcanzaron la mayor exactitud.

Por su parte, Sharma et al. (2023) ofrecen una visión general de la minería de datos educativos y del análisis del aprendizaje, y describen técnicas como el agrupamiento, los árboles de decisión y las reglas de asociación para identificar patrones ocultos en el rendimiento académico y en la retención en las instituciones de educación superior.

De acuerdo con la revisión sistemática de Colpo et al. (2024), los estudios sobre minería de datos educativos aplicada a la predicción de la deserción privilegian mayoritariamente el uso de datos académicos, demográficos y económicos de estudiantes de pregrado, combinados con técnicas de clasificación por conjuntos (*ensemble*) de árboles de decisión; los mismos autores advierten, además, que los hallazgos predictivos generados por estos modelos permanecen subutilizados en la práctica educativa y que la evidencia disponible proviene mayoritariamente de países desarrollados, lo que confirma que deben considerarse múltiples variables y contextos regionales para el diseño de estrategias de retención.

En relación con este planteamiento, Rodríguez et al. (2023) sostienen que la eficiencia de la predicción de la deserción se fortalece mediante el umbral de probabilidad optimizado en los modelos de aprendizaje automático. En el sector de la educación a distancia, las posibilidades aumentan cuando se emplean estas herramientas de pronóstico, al disminuir los falsos negativos; por ello, es posible crear tácticas puntuales y oportunas para mantener un flujo sólido de estudiantes.

En la misma línea, Ifenthaler y Yau (2020), en su revisión sistemática de estudios sobre analítica del aprendizaje, reportan enfoques que apoyan el éxito académico y la identificación de estudiantes en riesgo de abandono en la educación superior, y conciben esta analítica como una práctica sociotécnica para la cual advierten la falta de evidencia rigurosa a gran escala.

La utilización de un marco integral de aprendizaje automático facilita la realización de pronósticos precisos sobre el promedio académico de los estudiantes. En el estudio de Gul et al. (2025) se plantea que este tipo de marco posibilita realizar pronósticos muy precisos sobre el rendimiento estudiantil, para lo cual se consideran tanto factores demográficos (como la raza, el género, el nivel educativo parental o las condiciones económicas y sociales) como académicos (las calificaciones en lectura, escritura y matemáticas). Al analizar ocho algoritmos avanzados de aprendizaje automático (CatBoost, Gradient Boosting, AdaBoost, XGBoost, Lasso, K-Nearest Neighbors, Random Forest y Decision Trees), fue CatBoost el que alcanzó la mayor exactitud (87,46%), superando a Decision Trees y a Gradient Boosting.

En el estudio de Le y Nguyen (2024), la regresión logística superó a las máquinas de vectores de soporte, a los árboles de decisión y a Random Forest en la predicción del estado de alerta académica de estudiantes de informática, lo que favorece la identificación temprana de los estudiantes con riesgo de deserción. El valor de estas estrategias radica en que permiten prever a los estudiantes eventualmente en riesgo y brindan apoyo para el diseño de programas de retención más efectivos y sostenibles.

En este sentido, Contreras (2021) analizó, mediante árboles de decisión y técnicas de agrupamiento, los datos institucionales de una muestra de 250 estudiantes beneficiarios de una beca de nivelación académica en una universidad pública de Chile, y concluyó que el puntaje obtenido en la prueba específica de matemáticas de la selección universitaria permite discriminar el riesgo de deserción inicial, lo cual sustenta un sistema de alerta temprana para el acompañamiento focalizado de los estudiantes. Este hallazgo indica que los puntajes de admisión son un elemento al que debe prestarse especial atención en el diseño de estrategias de retención.

Revista Scientific - Artículo de Revisión Sistemática - Arbitrado - Volumen 11, Número 40 - Mayo-Julio 2026 - pp. 93-111
pp. BA2016000002 - e-ISSN: 2542-2987 - ISNI: 0000 0004 6045 0361

DOI: https://doi.org/10.29394/Scientific.issn.2542-2987.2026.11.40.6.93-111
OAI-PMH: http://www.indteca.com/ojs/index.php/Revista_Scientific/oai

1.3. Innovación Tecnológica, Gobernanza Universitaria y Estrategias de Retención

Más allá de los aspectos técnicos y metodológicos de los modelos predictivos, es necesario considerar el papel de la innovación tecnológica y de la gobernanza universitaria en el diseño de estrategias de retención sostenibles.

En el trabajo de Ismaili y Besimi (2024) se examina el desarrollo de un almacén de datos institucional que integra datos demográficos, de desempeño académico, calificaciones, asistencia y participación de los estudiantes para identificar, mediante aprendizaje automático, a quienes se encuentran en riesgo de fracaso académico, con el fin de apoyar la toma de decisiones institucionales y mejorar la retención y el éxito académico.

A partir de lo anterior, puede plantearse, como reflexión de los autores del presente artículo, que la incorporación de las TIC en la gestión académica requiere marcos de gobernanza universitaria que garanticen la responsabilidad, la transparencia, la calidad y la inclusión en el uso de los datos y de los modelos predictivos.

Por otro lado, Shafiq et al. (2022), en una revisión sistemática de la literatura, examinan los factores subyacentes y los aspectos metodológicos de la construcción de modelos predictivos para la retención estudiantil mediante la minería de datos educativos y la analítica del aprendizaje.

De forma complementaria, la evidencia sobre abandono, retención y analítica educativa muestra que los modelos predictivos deben integrar variables académicas, temporales, socioeconómicas y de interacción, y que su implementación requiere acompañamiento institucional (Choi, 2018; Cholis & Ulinnuha, 2023; Colpo et al., 2024; Ifenthaler & Yau, 2020; Mduma et al., 2019; Murnawan et al., 2024; Nicoletti & de Oliveira, 2020; Ouyang et al., 2023; Patacsil, 2020; Perkash et al., 2024; Shafiq et al., 2022; Tao, 2024; Xu et al., 2022; Ye & Biswas, 2014; Yükseltürk et al., 2014). Asimismo, los factores institucionales y de diseño del aprendizaje, como la flexibilidad, se relacionan con la persistencia y el abandono (Xavier & Meneses, 2021), y la salud mental se asocia con la probabilidad de abandonar los estudios (Zając et al., 2024).

Sin embargo, las revisiones previas sobre el tema tienen un alcance distinto al que aquí se aborda. Mduma et al. (2019) ofrecen un panorama general de los enfoques de aprendizaje automático para predecir la deserción escolar y advierten que la mayoría de los algoritmos se desarrollaron y probaron en países desarrollados; Shafiq et al. (2022) examinan, mediante una revisión sistemática, los factores asociados a la retención estudiantil y los aspectos metodológicos para construir modelos predictivos con minería de datos educativos y analítica del aprendizaje; Ifenthaler y Yau (2020) examinan la analítica del aprendizaje orientada al éxito académico; Murnawan et al. (2024) analizan los factores empleados para predecir el rendimiento, no la deserción; Colpo et al. (2024) advierten que la evidencia proviene mayoritariamente de países desarrollados, y Alves et al. (2024) abordan la deserción en una revisión de alcance de carácter general.

Ninguna de ellas reúne en un mismo análisis los algoritmos implementables en Python, su exactitud reportada y la producción latinoamericana publicada en español, incluidas las tesis de posgrado, lo que justifica una revisión sistemática actualizada hasta 2025 con ese enfoque.

A partir de lo expuesto, surge la siguiente pregunta de investigación: ¿qué modelos predictivos de aprendizaje automático, implementables en Python, y con qué exactitud reportada, se han empleado en la literatura científica para anticipar la deserción universitaria?. En correspondencia con esta interrogante, el presente trabajo tiene como objetivo principal sistematizar la evidencia científica disponible sobre el uso de modelos predictivos de aprendizaje automático, implementables en Python, para anticipar la deserción universitaria, mediante una revisión sistemática desarrollada conforme a las directrices

Revista Scientific - Artículo de Revisión Sistemática - Arbitrado - Volumen 11, Número 40 - Mayo-Julio 2026 - pp. 93-111
pp. BA2016000002 - e-ISSN: 2542-2987 - ISNI: 0000 0004 6045 0361

DOI: https://doi.org/10.29394/Scientific.issn.2542-2987.2026.11.40.6.93-111
OAI-PMH: http://www.indteca.com/ojs/index.php/Revista_Scientific/oai

PRISMA 2020 (Page et al., 2021).

2. Metodología

El presente estudio se correspondió, por su finalidad, con una investigación de tipo básica, orientada a sistematizar y generar conocimiento teórico en torno al fenómeno de la deserción universitaria y a su abordaje mediante modelos predictivos, sin perseguir una aplicación inmediata sobre una población determinada. En cuanto al enfoque, el estudio se sustentó en una perspectiva cualitativa, en la medida en que privilegió la síntesis narrativa y la interpretación crítica de fuentes documentales por sobre la medición estadística de variables (Hernández-Sampieri & Mendoza, 2018).

Respecto al método, se empleó el método analítico-sintético, el cual permitió descomponer los contenidos de las fuentes en sus elementos constitutivos (algoritmo, muestra, métrica de desempeño) para luego integrarlos en una síntesis interpretativa coherente, en concordancia con los principios generales del análisis documental (Cohen et al., 2018). Finalmente, en lo que concierne al diseño, el estudio se ajustó a un diseño de revisión sistemática, no experimental, en tanto no implicó la manipulación de variables ni la intervención directa sobre sujetos de estudio, sino el análisis de información ya existente en la literatura especializada (Creswell & Creswell, 2018).

El diseño se fundamentó, además, en las pautas para revisiones sistemáticas de Kitchenham y Charters (2007), concebidas para las áreas de ingeniería y computación, y en el Manual Cochrane (Higgins et al., 2024) para la formulación de la pregunta, la selección de los estudios y la evaluación de su calidad. Estos cuatro elementos, tipo, enfoque, método y diseño, se articularon de manera complementaria con las fases PRISMA 2020 descritas a continuación.

2.1. Protocolo y Registro

El protocolo de la presente revisión no fue registrado previamente en una plataforma de registro de protocolos, lo cual se reconoce como una limitación del proceso (véase el apartado de Discusión). No obstante, la revisión se informa conforme a la declaración PRISMA 2020 (Page et al., 2021), con el propósito de garantizar la transparencia y la replicabilidad del proceso de búsqueda, selección y extracción de datos.

2.2. Criterios de Elegibilidad

La pregunta de investigación se formuló bajo la estructura PICO/PEO adaptada a una revisión de tecnología educativa: Población (estudiantes de educación superior universitaria), Exposición / Intervención (modelos predictivos basados en técnicas de aprendizaje automático o minería de datos, implementables con herramientas de Python), Comparación (entre distintos algoritmos y contextos institucionales) y Resultado (exactitud o precisión predictiva reportada para anticipar la deserción). De ello se derivó la pregunta de investigación formulada al final de la Introducción.

Se incluyeron estudios aplicados en instituciones de educación superior, publicados entre 2014 y 2025, redactados en español o en inglés, que reportaran una técnica de aprendizaje automático o minería de datos identificable con fines predictivos de la deserción o el abandono estudiantil, y que ofrecieran una descripción metodológica suficiente para extraer al menos el algoritmo empleado y, cuando estuviera disponible, el tamaño de la muestra y una métrica de desempeño. Se consideraron elegibles artículos de revista, ponencias arbitradas y tesis con datos aplicados.

Se excluyeron las notas de prensa, entradas de blog, repositorios de código sin artículo académico asociado, conjuntos de datos sin estudio, *preprints* sin revisión por pares, estudios centrados en niveles educativos distintos al universitario, estudios sin foco explícito en la predicción de la deserción, revisiones o piezas teóricas secundarias (estas se citan en la Introducción y la Discusión, pero no se contabilizan como estudio primario incluido en la síntesis), publicaciones anteriores a 2014, y estudios cuyo texto completo no pudo recuperarse para verificar de forma fiable los datos de interés.

DOI: https://doi.org/10.29394/Scientific.issn.2542-2987.2026.11.40.6.93-111
OAI-PMH: http://www.indteca.com/ojs/index.php/Revista_Scientific/oai

2.3. Fuentes de Información

La búsqueda se ejecutó el 17 de julio de 2025 en Google Académico y en el buscador de SciELO, ambas fuentes de consulta gratuita. No se tuvo acceso a bases de datos de suscripción (Scopus, IEEE Xplore, SpringerLink), lo cual constituye una limitación reconocida del proceso (véase Discusión) frente al estándar deseable de cobertura multibase de una revisión sistemática.

2.4. Estrategia de Búsqueda

Se emplearon ocho cadenas de búsqueda, combinando términos relacionados con la deserción universitaria ("deserción universitaria", "abandono universitario", "*student dropout*", "*university dropout*") con términos relacionados con modelos predictivos y aprendizaje automático ("modelo predictivo", "*machine learning*", "Python", "Random Forest", "*educational data mining*", "Scikit-learn"), en las dos fuentes indicadas en la sección Fuentes de Información (2.3) y con un rango temporal de 2014 a 2025. Las ocho cadenas se derivaron de la siguiente ecuación general, adaptada a la sintaxis de cada fuente: ("deserción universitaria" OR "abandono universitario" OR "*student dropout*" OR "*university dropout*") AND ("modelo predictivo" OR "*machine learning*" OR "*educational data mining*" OR "Random Forest" OR "Python" OR "Scikit-learn").

2.5. Proceso de Selección

El proceso de selección fue realizado por los tres autores, sin apoyo de herramientas de inteligencia artificial. Cada registro fue evaluado de forma independiente por cada uno de ellos, y las discrepancias se resolvieron mediante discusión hasta alcanzar consenso. De los 80 registros identificados, se eliminaron 11 duplicados exactos por URL y se colapsaron 14 registros adicionales correspondientes a espejos del mismo estudio republicados en distintos repositorios (ResearchGate, Dialnet, academia.edu), resultando en 55 registros distintos. De estos, 1 ya se encontraba entre las 34 referencias de base seleccionadas por los autores antes de la búsqueda.

Los 54 registros restantes se cribaron por título y resumen aplicando los criterios de elegibilidad, de los cuales 22 fueron excluidos con motivo documentado. De los 32 informes buscados para su recuperación a texto completo, 12 no pudieron recuperarse por restricciones de acceso en repositorios de suscripción sin alternativa de acceso abierto disponible; los 20 restantes se evaluaron a texto completo, y 4 de ellos se excluyeron por tratarse de revisiones o piezas teóricas secundarias en lugar de estudios primarios, quedando finalmente 16 estudios nuevos incluidos en la síntesis. El diagrama de flujo PRISMA 2020 de este proceso se presenta en la figura 1 (apartado 3.1, Selección de Estudios).

2.6. Extracción de Datos

De cada estudio incluido se extrajeron los siguientes datos: autor(es) y año de publicación, país e institución donde se aplicó el estudio, algoritmo o algoritmos de aprendizaje automático empleados, tamaño de la muestra, y la métrica de desempeño reportada (exactitud, área bajo la curva ROC o *F1-score*, según lo que cada estudio hubiera reportado explícitamente), mediante la lectura directa del texto completo de cada estudio. La extracción fue realizada por los tres autores mediante una matriz común, y los datos fueron contrastados entre ellos para resolver por consenso cualquier discrepancia.

2.7. Evaluación de Calidad y Riesgo de Sesgo

Dada la heterogeneidad de diseños entre los estudios incluidos (artículos de revista, ponencias arbitradas y tesis), se aplicó una lista de verificación ad hoc de cinco criterios, evaluados de forma binaria (cumple/no cumple) para cada estudio: (1) objetivo de investigación claramente definido, (2) fuente y tamaño de la muestra descritos con precisión, (3) algoritmo o algoritmos de aprendizaje automático especificados de forma explícita, (4) métrica de desempeño reportada con una cifra concreta y verificable, y (5) procedimiento de validación (validación cruzada o partición entrenamiento / prueba) mencionado explícitamente. Esta evaluación fue aplicada por los tres autores, y los desacuerdos se

DOI: https://doi.org/10.29394/Scientific.issn.2542-2987.2026.11.40.6.93-111
OAI-PMH: http://www.indteca.com/ojs/index.php/Revista_Scientific/oai

resolvieron por consenso.

Se optó por esta lista ad hoc en lugar de una herramienta estandarizada como PROBAST *(Prediction model Risk Of Bias ASsessment Tool;* Wolff et al., 2019), diseñada específicamente para estudios de modelos de predicción, debido a que varios de los estudios incluidos, en particular las tesis, no reportaban la información necesaria para completar todos los dominios que dicha herramienta exige; esta limitación se reconoce explícitamente en el apartado de Discusión. La tabla 2 presenta los resultados de esta evaluación.

2.8. Métodos de Síntesis

Dada la heterogeneidad de algoritmos, contextos institucionales y métricas de desempeño reportadas por los estudios incluidos (exactitud, AUC y *F1-score*, no siempre convertibles entre sí de forma directa), no fue posible realizar un metaanálisis cuantitativo de los resultados. En su lugar, se optó por una síntesis narrativa, agrupando los hallazgos por ejes temáticos (familias de algoritmos, exactitud reportada, distribución geográfica y calidad metodológica de los estudios), en línea con el enfoque cualitativo-documental declarado para el conjunto del estudio.

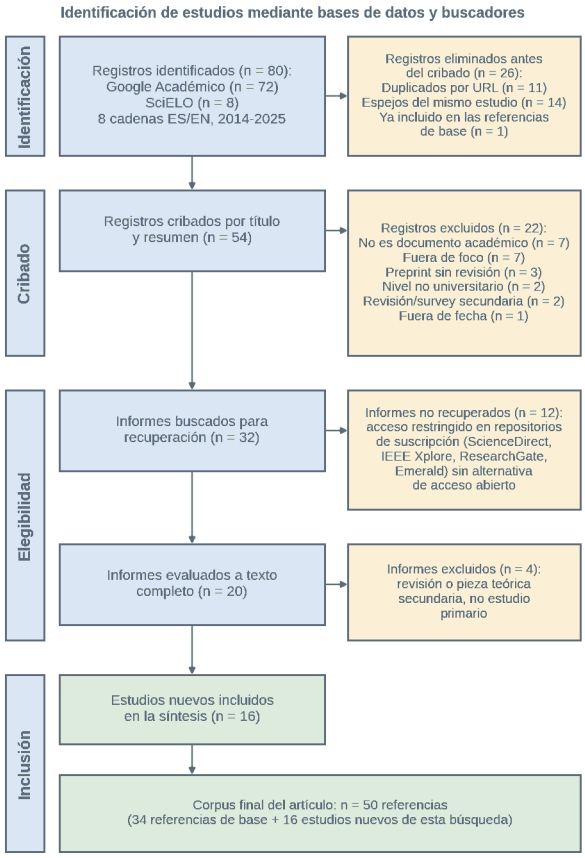
3. Resultados

Esta sección se organiza en cuatro apartados que siguen la secuencia del proceso metodológico descrito en la sección anterior: primero se expone la selección de los estudios; a continuación, se describen sus características principales; luego se presenta la evaluación de su calidad metodológica, y, por último, se ofrece la síntesis narrativa de los hallazgos, organizada según los ejes temáticos definidos en los Métodos de Síntesis (2.8).

3.1. Selección de Estudios

El proceso de búsqueda, cribado y selección de estudios se desarrolló conforme a lo descrito en el apartado de Metodología, siguiendo las directrices PRISMA 2020 (Page et al., 2021). La Figura 1 presenta el diagrama de flujo correspondiente a este proceso.

Figura 1. Diagrama de Flujo PRISMA 2020 del Proceso de Búsqueda y Selección de Estudios.



Nota. Fuente: Zambrano et al. (2025), siguiendo las directrices PRISMA 2020 (Page et al., 2021). Fuentes de consulta gratuita (Google Académico y SciELO); no se tuvo acceso a bases de datos de suscripción.

Como se observa en la figura 1, de los 80 registros identificados mediante la búsqueda en Google Académico y SciELO, y tras la eliminación de duplicados y espejos del mismo estudio, el cribado por título y resumen, y la evaluación a texto completo, se incluyeron 16 estudios nuevos en la síntesis, los cuales se sumaron a 34 referencias de base, seleccionadas por los autores antes de la búsqueda para el marco conceptual y metodológico, para conformar un corpus final de 50 referencias temáticas; las restantes entradas de la lista de Referencias corresponden a obras metodológicas que fundamentan el diseño de la revisión.

La proporción relativamente baja de estudios finalmente incluidos (20% de los registros identificados) se explica, según la Figura 1, por la elevada proporción

Revista Scientific - Artículo de Revisión Sistemática - Arbitrado - Volumen 11, Número 40 - Mayo-Julio 2026 - pp. 93-111
pp. BA2016000002 - e-ISSN: 2542-2987 - ISNI: 0000 0004 6045 0361

DOI: https://doi.org/10.29394/Scientific.issn.2542-2987.2026.11.40.6.93-111
OAI-PMH: http://www.indteca.com/ojs/index.php/Revista_Scientific/oai

de duplicados y espejos del mismo estudio en distintos repositorios (25 registros), por las exclusiones en el cribado por título y resumen (22 registros), por la imposibilidad de recuperar el texto completo de estudios alojados en repositorios de suscripción sin alternativa de acceso abierto (12 informes) y, en menor medida, por la exclusión de revisiones o piezas teóricas secundarias sin datos primarios propios (4 informes). Las 34 referencias de base no fueron sometidas al proceso de selección descrito ni a la evaluación de calidad de la sección de Evaluación de Calidad y Riesgo de Sesgo (2.7), por lo que las conclusiones sobre algoritmos, exactitud y calidad metodológica se sustentan únicamente en los 16 estudios nuevos incluidos.

3.2. Características de los Estudios Incluidos

Los 16 estudios nuevos incluidos en la síntesis proceden de autores de nueve países de cuatro continentes, aunque seis de ellos analizan un mismo conjunto de datos público de Portugal, con tamaños de muestra que oscilan entre 104 y 32.593 registros de estudiantes. La tabla 1 sintetiza las características principales de cada estudio: autores y año, país o contexto institucional, algoritmo(s) empleado(s), tamaño de muestra y métrica de desempeño reportada.

Tabla 1. Características de los Estudios Nuevos Incluidos en la Síntesis.

Autor(es) y año	País de los autores / origen de los datos	Algoritmo(s)	Muestra	Métrica de desempeño
Guadalupe (2025)	Perú	Random Forest (único algoritmo)	104 estudiantes (población de 143)	Random Forest: Exactitud 96,87%
Zárate-Valderrama et al. (2021)	Perú (Universidad Nacional de San Agustín)	Red Neuronal, ID3, C4.5	300 registros útiles (de 970)	C4.5: Exactitud 68% (mejor algoritmo)
Polo (2024)	Perú	Ensamble por votación (Bayesiano, Regresión, SVM, Random Forest)	500 registros (2012-2021; partición 80/20)	Exactitud global: 93%
Aco et al. (2023)	Perú / Portugal	Regresión Logística, Naive Bayes, MLP, Árbol de Decisión, SVM, Random Forest	2.422 registros	Regresión Logística: 89,84% (mejor algoritmo)
Ramírez y Grandón (2018)	Chile	Árbol de decisión con parámetros optimizados (CBAD)	5.288 estudiantes	Razón de precisión: 87,27%
Quintana (2023)	Perú (Universidad Nacional de Moquegua)	Árbol de Decisión, Regresión Logística, SVM, Naive Bayes	329 registros (109 desertores y 220 no desertores) de 2.305	Árbol de Decisión: 80,22% con 12 características (mejor algoritmo)
Marcolino et al. (2025)	Brasil (Moodle, UFSC y UFPEL)	CatBoost (comparado con Random Forest, XGBoost, k-NN, Regresión Logística, Naive Bayes)	567 instancias estudiante-curso (23 cursos)	Exactitud: 90%; AUC: 0,95 (hold-out 30%)
Al Hashmi et al. (2025)	Emiratos Árabes Unidos	XGBoost, GBM, Random Forest, CART, Regresión Logística, KNN	6.798 estudiantes	AUC-ROC superior a 0,91 (XGBoost, GBM y Random Forest)
Winarsih et al. (2025)	Indonesia (conjunto de datos público OULAD, Reino Unido)	Árbol de Decisión, Naive Bayes, Regresión Logística, Random Forest, AdaBoost	OULAD completo (32.593 estudiantes)	Random Forest: Exactitud 89,37% (partición 80:20); F1 94,45% (validación cruzada)
Alameri (2025)	Europa (institución no identificada; 4.424 registros coincidentes con el conjunto público de Portugal)	Random Forest, XGBoost, ensamble por votación suave	4.424 registros (35 variables, 10 años)	XGBoost ajustado: exactitud balanceada 91,5%; AUC 0,957; F1 92,1% (validación cruzada de 5 pliegues)
Putra et al. (2025)	Indonesia	Random Forest, XGBoost	4.424 instancias	Random Forest: 80,56%
Olive et al. (2025)	Ruanda	Regresión Logística, Random Forest, AdaBoost (votación suave y dura)	4.424 registros (3.539 analizados)	Votación suave: Exactitud 80,56%; AUC: 0,91
Osemwegie y Amadin (2023)	Nigeria (Universidad de Benin)	Naive Bayes, Regresión Logística, SVM, Árbol de Decisión, KNN, Red Neuronal	906 registros	Regresión Logística: 98,9% (mejor algoritmo)
Villar y de Andrade (2024)	Brasil / Portugal	Árbol de Decisión, SVM, Random Forest, Gradient Boosting, XGBoost, CatBoost, LightGBM	4.424 estudiantes (datos de origen portugués; 1.325 instancias en el conjunto de prueba con SMOTE)	LightGBM y CatBoost: F1 86%-88%
Romero y Liao (2025)	Portugal	Lasso, GAM, Random Forest, XGBoost, Red Neuronal	4.424 → 3.400 registros	XGBoost: Exactitud 88,2%; F1 90,4%
Sulak y Koklu (2024)	Turquía	Árbol de Decisión, Random Forest, Red Neuronal Artificial	4.424 registros	Red Neuronal Artificial: Exactitud 77,3% (mejor algoritmo)

Nota. Fuente: Zambrano et al. (2025). Seis estudios (Alameri, 2025; Olive et al., 2025; Putra et al., 2025; Romero y Liao, 2025; Sulak y Koklu, 2024; Villar y de Andrade, 2024) analizan el mismo conjunto de datos público de 4.424 estudiantes de una institución de educación superior de Portugal, a partir de los 16 estudios nuevos incluidos tras el proceso de búsqueda y cribado PRISMA 2020 descrito en la Metodología.

Como se observa en la tabla 1, los estudios incluidos emplearon predominantemente algoritmos de ensamble (Random Forest, XGBoost, CatBoost, LightGBM) y, en menor medida, algoritmos lineales (regresión logística) y de árbol simple (árbol de decisión), aplicados sobre muestras que en su mayoría superaron los 1.000 registros. Dos estudios evaluaron un único algoritmo: Guadalupe (2025), con Random Forest, y Ramírez y Grandón (2018), con un árbol de decisión C4.5 optimizado; por ello no permiten comparar algoritmos.

Asimismo, Marcolino et al. (2025) modelan la reprobación o el riesgo académico en cursos en línea, un

Revista Scientific - Artículo de Revisión Sistemática - Arbitrado - Volumen 11, Número 40 - Mayo-Julio 2026 - pp. 93-111
pp. BA2016000002 - e-ISSN: 2542-2987 - ISNI: 0000 0004 6045 0361

DOI: https://doi.org/10.29394/Scientific.issn.2542-2987.2026.11.40.6.93-111
OAI-PMH: http://www.indteca.com/ojs/index.php/Revista_Scientific/oai

desenlace próximo a la deserción, pero no idéntico. La extracción de datos no registró de forma sistemática el lenguaje de programación ni las bibliotecas empleadas en cada estudio; por ello, la síntesis describe los algoritmos y no puede afirmar cuáles de los estudios los implementaron en Python, aunque las técnicas identificadas son de uso corriente en bibliotecas de código abierto de este lenguaje, como *Scikit-learn.*

3.3. Evaluación de Calidad

Se aplicó a cada uno de los 16 estudios incluidos la lista de verificación de cinco criterios descrita en el apartado de Metodología (2.7). La tabla 2 presenta el puntaje obtenido por cada estudio.

Tabla 2. Evaluación de Calidad Metodológica de los Estudios Nuevos Incluidos.

Autor(es) y año	Puntaje (sobre 5 criterios)
Guadalupe (2025)	5/5
Zárate-Valderrama et al. (2021)	5/5
Polo (2024)	5/5
Aco et al. (2023)	5/5
Ramírez y Grandón (2018)	5/5
Quintana (2023)	5/5
Marcolino et al. (2025)	5/5
Al Hashmi et al. (2025)	5/5
Winarsih et al. (2025)	5/5
Alameri (2025)	5/5
Putra et al. (2025)	5/5
Olive et al. (2025)	4/5
Osemwegie y Amadin (2023)	5/5
Villar y de Andrade (2024)	5/5
Romero y Liao (2025)	5/5
Sulak y Koklu (2024)	5/5

Nota. Fuente: Zambrano et al. (2025). Criterios: (1) objetivo definido, (2) muestra descrita, (3) algoritmo(s) especificado(s), (4) métrica con cifra concreta, (5) validación mencionada explícitamente. Promedio del conjunto: 4,9/5.

Como se observa en la tabla 2, el promedio de calidad metodológica del conjunto de estudios incluidos fue de 4,9 sobre 5 puntos posibles, lo cual indica un nivel de detalle metodológico alto en las fuentes, valorado a partir de la lectura del texto completo de cada estudio. Quince de los dieciséis estudios cumplieron los cinco criterios en su totalidad; la única excepción fue Olive et al. (2025), con 4/5, porque su texto no menciona de forma explícita un procedimiento de validación cruzada ni una partición entrenamiento/prueba. Este resultado describe la completitud del reporte y no equivale a una ausencia de riesgo de sesgo, dado que la lista de verificación es ad hoc y binaria (véase la sección de Discusión).

3.4. Síntesis de Resultados

La síntesis narrativa de los 16 estudios nuevos, integrada con la literatura de base expuesta en la Introducción, es coherente con la concepción de la deserción universitaria como un fenómeno multidimensional en el que convergen factores académicos, personales, socioeconómicos e institucionales.

En primer lugar, respecto a las familias de algoritmos empleadas, los estudios revisados evidenciaron una tendencia hacia los métodos de ensamble: Random Forest fue el algoritmo más frecuente entre los estudios incluidos (12 de 16), seguido de XGBoost (6 de 16) y, en los estudios más recientes (2024-2025), de CatBoost y LightGBM.

Las comparaciones directas con la regresión logística arrojaron resultados heterogéneos: en Osemwegie y Amadin (2023) la regresión logística obtuvo el mejor desempeño (98,9%) entre seis clasificadores; en Aco et al. (2023) fue también la regresión logística la que alcanzó el mejor desempeño relativo (89,84%), por encima de Random Forest y del resto de algoritmos comparados; y, en cambio, en Olive et al. (2025) el ensamble por votación suave (exactitud de 80,56%) superó a los clasificadores individuales, entre ellos la regresión logística y Random Forest. Esta heterogeneidad matiza la narrativa dominante sobre una superioridad universal de los métodos de ensamble y sugiere que su ventaja depende en buena medida del contexto institucional y de la calidad de los datos disponibles.

En segundo lugar, en cuanto a la exactitud reportada, esta osciló entre 68% (Zárate-Valderrama et al., 2021, Universidad Nacional de San Agustín, Perú, mediante C4.5) y 98,9% (Osemwegie & Amadin, 2023, Universidad de Benin, Nigeria, mediante regresión logística), con valores intermedios como el 93% de exactitud global reportado por Polo (2024); los estudios

Revista Scientific - Artículo de Revisión Sistemática - Arbitrado - Volumen 11, Número 40 - Mayo-Julio 2026 - pp. 93-111
pp. BA2016000002 - e-ISSN: 2542-2987 - ISNI: 0000 0004 6045 0361

DOI: https://doi.org/10.29394/Scientific.issn.2542-2987.2026.11.40.6.93-111
OAI-PMH: http://www.indteca.com/ojs/index.php/Revista_Scientific/oai

que reportaron el área bajo la curva ROC alcanzaron valores de hasta 0,95 en Marcolino et al. (2025), en Brasil, mediante CatBoost, y hasta 0,957 en Alameri (2025). Esta distinción entre métricas, no siempre convertibles entre sí, evidencia una amplia heterogeneidad asociada al tamaño de muestra, la calidad de los datos disponibles y el contexto institucional.

Por su parte, Marcolino et al. (2025) combinaron algoritmos con técnicas de balanceo de datos (ADASYN) para atender la naturaleza desbalanceada de la deserción como variable de clasificación, mientras que Romero y Liao (2025) incorporaron un componente cuasiexperimental para evaluar el efecto de las becas sobre la probabilidad de deserción y Al Hashmi et al. (2025) aplicaron técnicas de explicabilidad para interpretar los predictores del abandono.

En contraste, las tesis doctorales peruanas de Guadalupe (2025), Polo (2024) y Quintana (2023) aportaron evidencia aplicada a instituciones específicas del país, con distinto alcance comparativo: mientras Polo (2024) reportó una exactitud global del 93% para un modelo de ensamble por votación, Guadalupe (2025) reportó una exactitud de 96,87% con Random Forest, único algoritmo evaluado en su estudio.

De igual modo, Quintana (2023) identificó al árbol de decisión (80,22% con las 12 características seleccionadas mediante Random Forest) como el algoritmo de mejor desempeño relativo en su estudio comparativo, mientras que en Alameri (2025) el modelo XGBoost ajustado superó al ensamble por votación suave.

En tercer lugar, según la afiliación de los autores, América Latina aportó la mitad de los estudios incluidos (8 de 16: cinco de Perú, uno de Chile y dos de Brasil), con predominio de tesis en el caso peruano, seguida de Asia (Emiratos Árabes Unidos, Indonesia y Turquía), África (Ruanda y Nigeria) y Europa (Portugal). Según el origen de los datos, en cambio, seis estudios analizaron el mismo conjunto público de 4.424 estudiantes de una institución portuguesa y Winarsih et al. (2025) empleó el conjunto OULAD, del Reino Unido, de modo que solo nueve estudios trabajaron con datos de su propia institución o país. Ningún estudio incluido corresponde a Ecuador.

Por último, la evaluación de calidad metodológica (tabla 2) mostró un reporte completo en casi la totalidad de los estudios, lo cual es coherente con la relación, señalada en la Introducción, entre la transparencia metodológica y la utilidad práctica de los modelos predictivos como insumo para el diseño de políticas de retención estudiantil.

4. Discusión

Los resultados presentados, leídos junto con la literatura de base, son coherentes con la idea de que la deserción en la educación superior difícilmente puede evaluarse y comprenderse de manera unidimensional y que requiere integrar perspectivas que atiendan de manera simultánea sus dimensiones académica, socioeconómica, personal e institucional. Esta lectura coincide con lo indicado por Munizaga et al. (2018) y Ramírez y Carcausto-Calla (2024), quienes sostienen que la deserción estudiantil es un fenómeno multicausal, con impacto en el estudiante, la familia y el Estado.

Entre los estudios incluidos, los métodos de ensamble fueron los más empleados y alcanzaron un desempeño elevado en varios contextos, aunque la regresión logística los superó en Osemwegie y Amadin (2023) y en Aco et al. (2023). Este patrón es parcialmente coherente con la literatura previa: Gul et al. (2025) reportan la superioridad de CatBoost en la predicción del rendimiento, mientras que Urbina-Nájera et al. (2021) destacan la interpretabilidad de los árboles de decisión para identificar a los estudiantes en riesgo.

La literatura previa sugiere, además, que la combinación de variables académicas, demográficas y de interacción en ambientes de aprendizaje virtual aumenta la capacidad predictiva (Perkash et al., 2024) y que la predicción resulta más útil cuando orienta acciones de acompañamiento (Ouyang et al., 2023). En este punto se evidencia un reto metodológico relevante,

Revista Scientific - Artículo de Revisión Sistemática - Arbitrado - Volumen 11, Número 40 - Mayo-Julio 2026 - pp. 93-111
pp. BA2016000002 - e-ISSN: 2542-2987 - ISNI: 0000 0004 6045 0361

DOI: https://doi.org/10.29394/Scientific.issn.2542-2987.2026.11.40.6.93-111
OAI-PMH: http://www.indteca.com/ojs/index.php/Revista_Scientific/oai

en tanto los modelos deben tener en cuenta las dinámicas temporales y contextuales que advierten que el riesgo de deserción es cambiante y se transforma a lo largo del ciclo educativo.

De igual manera, el debate actual se extiende más allá de lo técnico y se enfoca en la gestión que realizan las universidades. El reto radica en la conexión de las herramientas tecnológicas con los procesos de gestión institucional. Dado que los factores económicos e institucionales inciden en el abandono (Munizaga et al., 2018), la predicción anticipada corre el riesgo de convertirse en un simple diagnóstico si no se acompaña de programas de becas, soporte psicosocial y ayuda financiera.

En este sentido, los autores del presente estudio consideran que la articulación entre los modelos predictivos y las políticas institucionales resulta, en última instancia, más determinante para la retención estudiantil que el algoritmo específico empleado, en la medida en que la precisión técnica del modelo carece de efecto práctico si sus resultados no se traducen en decisiones institucionales oportunas y sostenidas en el tiempo. En consecuencia, la incorporación de bibliotecas como Scikit-learn, Pandas y NumPy en los procesos institucionales de analítica académica solo genera valor real cuando se articula con protocolos claros de gobernanza de datos y con la capacitación del personal encargado de interpretar sus resultados.

De esta manera, el debate internacional sobre la deserción universitaria concuerda en que deben implementarse enfoques integrales donde la tecnología sea el instrumento y no un fin en sí mismo. Para alcanzar métodos más humanos y efectivos de retención estudiantil, la asociación entre salud mental y abandono documentada por Zając et al. (2024) sugiere, como inferencia de los autores del presente artículo, considerar variables de bienestar en los modelos predictivos.

En síntesis, la evidencia revisada sugiere que la disminución de la deserción universitaria podría obtener mejores resultados mediante la combinación de estrategias multidimensionales, políticas institucionales integrales e instrumentos tecnológicos. El desafío futuro consiste en extender estas experiencias a nivel regional, particularmente en América Latina, donde los riesgos aumentan debido a los elementos estructurales propios de la región.

Los hallazgos de la presente revisión son consistentes con los reportados por Ifenthaler y Yau (2020), quienes reúnen variables, algoritmos y métodos de la analítica del aprendizaje aplicados al éxito académico y al riesgo de abandono, así como con Mduma et al. (2019) y Colpo et al. (2024), quienes advierten que la evidencia sobre la predicción de la deserción procede mayoritariamente de países desarrollados, lo que coincide con que siete de los estudios incluidos se apoyen en conjuntos de datos de instituciones europeas. La presencia de ocho estudios de autores latinoamericanos, entre ellos tres tesis doctorales peruanas (Guadalupe, 2025; Polo, 2024; Quintana, 2023), sugiere, no obstante, una creciente visibilidad de la producción regional en fuentes de acceso abierto, aunque concentrada en Perú y apoyada en parte en datos de otros países.

La presente revisión sistemática presenta limitaciones metodológicas que deben reconocerse explícitamente. En primer lugar, el protocolo no fue registrado previamente en una plataforma de registro de protocolos, como OSF Registries o INPLASY, lo cual reduce la transparencia a priori del proceso frente al estándar deseable de una revisión sistemática. En segundo lugar, la búsqueda se limitó a fuentes de consulta gratuita (Google Académico y SciELO), sin acceso a bases de datos de suscripción como Scopus, IEEE Xplore o SpringerLink, y a publicaciones en español e inglés, lo cual pudo excluir estudios relevantes alojados exclusivamente en dichas bases o publicados en otros idiomas; de hecho, 12 de los 32 informes buscados para su recuperación a texto completo no pudieron recuperarse precisamente por esta restricción de acceso.

En tercer lugar, la evaluación de calidad se realizó mediante una lista de verificación ad hoc de cinco

Revista Scientific - Artículo de Revisión Sistemática - Arbitrado - Volumen 11, Número 40 - Mayo-Julio 2026 - pp. 93-111
pp. BA2016000002 - e-ISSN: 2542-2987 - ISNI: 0000 0004 6045 0361

DOI: https://doi.org/10.29394/Scientific.issn.2542-2987.2026.11.40.6.93-111
OAI-PMH: http://www.indteca.com/ojs/index.php/Revista_Scientific/oai

criterios en lugar de una herramienta estandarizada como PROBAST, diseñada específicamente para estudios de modelos de predicción, debido a que varios estudios, en particular las tesis, no reportaban la información necesaria para completar todos los dominios que dicha herramienta exige. En cuarto lugar, la heterogeneidad de las métricas de desempeño reportadas por los estudios incluidos (exactitud, área bajo la curva ROC y *F1-score*, no siempre convertibles entre sí) impidió la realización de un metaanálisis cuantitativo, restringiendo la síntesis a un enfoque narrativo.

En quinto lugar, es necesario reconocer un posible sesgo de publicación: la totalidad de los 16 estudios nuevos incluidos reporta resultados de desempeño predictivo positivo (exactitudes entre 68% y 98,9% y valores de AUC de hasta 0,957), sin que se identificara ningún estudio con resultados predictivos débiles o negativos, lo cual es consistente con la conocida tendencia de la literatura científica a publicar predominantemente hallazgos favorables.

En sexto lugar, dado que Random Forest fue uno de los términos empleados en la estrategia de búsqueda (véase 2.4), su alta frecuencia entre los estudios incluidos (12 de los 16, 75%) no debe interpretarse como una medida no sesgada de su prevalencia real en el campo, sino, al menos en parte, como un artefacto de la propia estrategia de recuperación documental; a ello se suma que 6 de los 16 estudios nuevos (Alameri, 2025; Olive et al., 2025; Putra et al., 2025; Romero & Liao, 2025; Sulak & Koklu, 2024; Villar & de Andrade, 2024) informan una muestra de 4.424 registros que coincide con la del conjunto público de origen portugués, lo que reduce la independencia real de la evidencia sintetizada en esa proporción del corpus.

Por último, la extracción de datos no registró de forma sistemática el lenguaje de programación ni las bibliotecas empleadas, por lo que esta revisión no puede establecer cuántos de los estudios incluidos utilizaron Python. Asimismo, algunas fuentes presentan inconsistencias internas que se registraron sin corregir las cifras de los autores: Quintana (2023) describe su muestreo de 329 registros como probabilístico estratificado en el resumen y como no aleatorio en la metodología; Marcolino et al. (2025) informan 567 instancias estudiante-curso, pero sus subconjuntos de entrenamiento y prueba (220 y 357) suman 577; y Winarsih et al. (2025) presentan el valor F1 de *Random Forest* como 9,45% en el resumen y 94,45% en las conclusiones, por lo que se adoptó este último por ser coherente con las demás métricas del estudio. Además, en Winarsih et al. (2025) y Polo (2024) las métricas por algoritmo solo aparecen en figuras, por lo que la Tabla 1 recoge únicamente la cifra del mejor modelo.

Pese a estas limitaciones, la revisión cuenta con una fortaleza metodológica relevante: la búsqueda, el cribado, la selección, la extracción de datos y la evaluación de calidad fueron realizados por los tres autores, quienes evaluaron cada registro de forma independiente y resolvieron las discrepancias por consenso, sin apoyo de herramientas de inteligencia artificial. Este procedimiento, basado en la participación de más de un revisor, práctica cuyo reporte exige PRISMA 2020 (Page et al., 2021), reduce el riesgo de sesgo de selección y de errores en la extracción, y refuerza la fiabilidad de la síntesis presentada.

5. Conclusiones

En respuesta a la pregunta de investigación y al objetivo planteados, el estudio permite concluir que los modelos predictivos de aprendizaje automático, implementables en Python, se perfilan como herramientas técnicamente prometedoras para anticipar la deserción universitaria, en la medida en que posibilitan identificar con antelación a los estudiantes en situación de riesgo; la literatura de base sugiere, además, que integrar variables académicas, socioeconómicas, demográficas e institucionales favorece esa capacidad predictiva. Algoritmos como la regresión logística, el Random Forest y los árboles de decisión alcanzan, en varios de los estudios revisados, niveles de exactitud que superan el 80% e incluso el 90% en la predicción

Revista Scientific - Artículo de Revisión Sistemática - Arbitrado - Volumen 11, Número 40 - Mayo-Julio 2026 - pp. 93-111
pp. BA2016000002 - e-ISSN: 2542-2987 - ISNI: 0000 0004 6045 0361

DOI: https://doi.org/10.29394/Scientific.issn.2542-2987.2026.11.40.6.93-111
OAI-PMH: http://www.indteca.com/ojs/index.php/Revista_Scientific/oai

temprana del riesgo de abandono, aunque con notable heterogeneidad según el contexto institucional y el algoritmo empleado.

No obstante, la utilidad de estos instrumentos no se agota en su exactitud estadística, sino que exige un abordaje ético, comprensible y sustentado, capaz de traducirse en lineamientos institucionales para la toma de decisiones. En este sentido, la predicción resulta necesaria, pero no suficiente, pues su eficacia depende de que se articule con políticas institucionales orientadas a la equidad, el bienestar emocional, la inclusión y la sostenibilidad en el ámbito educativo.

Esta articulación resulta particularmente relevante en el contexto latinoamericano y, de manera prospectiva, en el ecuatoriano, para el cual no se identificaron estudios en esta revisión, donde la disponibilidad de herramientas de código abierto como Python reduce las barreras económicas para que instituciones con recursos limitados incorporen la analítica predictiva en su gestión académica.

Un aspecto novedoso de la presente revisión radica en la sistematización integradora de evidencia dispersa sobre modelos de aprendizaje automático implementables con bibliotecas de código abierto de Python como Scikit-learn, aplicados a la predicción de la deserción universitaria, con atención específica a la producción latinoamericana publicada en español. A diferencia de otras revisiones de carácter más general sobre analítica educativa, este estudio articula en un mismo cuerpo interpretativo los fundamentos teóricos de la deserción, los algoritmos de aprendizaje automático empleados y su desempeño en contextos institucionales concretos, lo cual aporta una mirada integral que combina distintos frentes de acción, desde la segmentación de grupos en riesgo hasta el diseño de prácticas de acompañamiento personalizadas y eficaces.

No obstante, es preciso reconocer las limitaciones de este estudio, detalladas en la Discusión: se trata de una revisión de carácter cualitativo, con síntesis narrativa, cuyas conclusiones dependen de los resultados reportados por terceros, y la heterogeneidad de los contextos institucionales y algorítmicos dificulta la comparación directa de las tasas de exactitud reportadas. Además, la restricción de la búsqueda a publicaciones en español e inglés pudo dejar fuera evidencia relevante disponible en otros idiomas.

De cara a futuras investigaciones, se recomienda que los estudios reporten de forma explícita el lenguaje de programación, las bibliotecas y las métricas de desempeño empleadas, a fin de facilitar la comparación entre modelos y la replicación de sus resultados. Asimismo, resulta pertinente que próximos estudios emprendan la validación empírica de los modelos predictivos aquí revisados mediante su aplicación directa con datos institucionales reales de universidades ecuatorianas, como la Universidad San Gregorio de Portoviejo, lo cual permitiría contrastar la evidencia internacional sistematizada en este trabajo con las particularidades del contexto local.

En definitiva, el desafío más significativo para las universidades de la región consiste en articular la innovación tecnológica con políticas públicas que atiendan los factores de riesgo social y económico asociados al abandono, de modo que la combinación de tecnología, compromiso social y gestión institucional se erige como una prioridad ineludible para la educación superior del siglo XXI.

6. Referencias

Aco, A. E., Hancco, B. O., & Pérez, Y. (2023). Análisis comparativo de técnicas de machine learning para la predicción de casos de deserción universitaria. *RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação*, (51), 84–98. https://doi.org/10.17013/risti.51.84-98

Al Hashmi, R. A. M., Zervopoulos, P. D., Elmehdi, H. M., & Ozturk, I. (2025). Predicting dropout in MENA STEM higher education using explainable AI: A machine learning approach. *Emerging Science Journal, 9*, 268–286. https://doi.org/10.28991/esj-2025-sied1-016

Alameri, F. (2025). *Predicting student dropout risk using machine learning* [Tesis de maestría, Rochester Institute of Technology]. RIT Scholar Works.

Alves, C., Queiroz, G., & Pilatti, L. A. (2024). Higher education dropout: A scoping review. *Revista de Gestão Social e Ambiental, 18*(8), Article e07156. https://doi.org/10.24857/rgsa.v18n8-117

Choi, Y. (2018). Student employment and persistence:

Revista Scientific - Artículo de Revisión Sistemática - Arbitrado - Volumen 11, Número 40 - Mayo-Julio 2026 - pp. 93-111
pp. BA2016000002 - e-ISSN: 2542-2987 - ISNI: 0000 0004 6045 0361

DOI: https://doi.org/10.29394/Scientific.issn.2542-2987.2026.11.40.6.93-111
OAI-PMH: http://www.indteca.com/ojs/index.php/Revista_Scientific/oai

Evidence of effect heterogeneity of student employment on college dropout. *Research in Higher Education, 59*(1), 88–107. https://doi.org/10.1007/s11162-017-9458-y

Cholis, D. N., & Ulinnuha, N. (2023). An ensemble voting approach for dropout student classification using decision tree C4.5, K-nearest neighbor and backpropagation. *Indonesian Journal of Artificial Intelligence and Data Mining, 6*(1), 107–115. https://doi.org/10.24014/ijaidm.v6i1.23412

Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education* (8th ed.). Routledge.

Colpo, M. P., Thompsen, T., Aguiar, M. S. D., & Cechinel, C. (2024). Educational data mining for dropout prediction: Trends, opportunities, and challenges. *Revista Brasileira de Informática na Educação, 32*, 220–256. https://doi.org/10.5753/rbie.2024.3559

Contreras, C. (2021). Determinación de variables predictivas de deserción inicial para generar un sistema de alerta temprana. Análisis sobre una muestra de estudiantes beneficiarios de la beca de nivelación académica en una universidad pública en Chile. *Calidad en la Educación*, (54), 12–45. https://doi.org/10.31619/caledu.n54.828

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE Publications.

Guadalupe, V. H. (2025). *Modelo predictivo basado en machine learning para la reducción de la deserción estudiantil en las universidades privadas del Perú: Caso Universidad Privada San Juan Bautista* [Tesis doctoral]. Universidad Nacional Federico Villarreal.

Gul, M. N., Abbasi, W., Babar, M. Z., Aljohani, A., & Arif, M. (2025). Data driven decisions in education using a comprehensive machine learning framework for student performance prediction. *Discover Computing, 28*(1), Article 153. https://doi.org/10.1007/s10791-025-09585-3

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

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2024). *Cochrane handbook for systematic reviews of interventions* (versión 6.5). Cochrane.

Ifenthaler, D., & Yau, J. Y. K. (2020). Utilising learning analytics to support study success in higher education: A systematic review. *Educational Technology Research and Development, 68*(4), 1961–1990. https://doi.org/10.1007/s11423-020-09788-z

Ismaili, B., & Besimi, A. (2024). A data warehousing framework for predictive analytics in higher education: A focus on student at-risk identification. *SEEU Review, 19*(2), 43–57. https://doi.org/10.2478/seeur-2024-0020

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering* (Technical Report EBSE-2007-01). Keele University; Durham University.

Le, M. T., & Nguyen, Q. A. (2024). Predicting learning status of informatics students at Dong Thap University based on machine learning models. *Dong Thap University Journal of Science - Natural Sciences Issue, 13*(5), 45–52. https://doi.org/10.52714/dthu.13.5.2024.1287

Marcolino, M. R., Porto, T. R., Primo, T. T., Targino, R., Ramos, V., Queiroga, E. M., … Cechinel, C. (2025). Student dropout prediction through machine learning optimization: Insights from moodle log data. *Scientific Reports, 15*(1), Article 9840. https://doi.org/10.1038/s41598-025-93918-1

Mduma, N., Kalegele, K., & Machuve, D. (2019). A survey of machine learning approaches and techniques for student dropout prediction. *Data Science Journal, 18*(1), Article 14. https://doi.org/10.5334/dsj-2019-014

Munizaga, F. R., Cifuentes, M. B., & Beltrán, A. J. (2018). Retención y abandono estudiantil en la educación superior universitaria en América Latina y el Caribe: Una revisión sistemática. *Education Policy Analysis Archives, 26*, Article 61. https://doi.org/10.14507/epaa.26.3348

Murnawan, M., Lestari, S., Samihardjo, R., & Dewi, D. A. (2024). Sustainable educational data mining studies: Identifying key factors and techniques for predicting student academic performance. *Journal of Applied Data Sciences, 5*(3), 1325–1342. https://doi.org/10.47738/jads.v5i3.347

Nicoletti, M. D. C., & de Oliveira, O. L. (2020). A machine learning-based computational system proposal aiming at higher education dropout prediction. *Higher Education Studies, 10*(4), 12. https://doi.org/10.5539/hes.v10n4p12

Olive, U., Bosco, M. J., & Enan, N. M. (2025). Predicting student dropout in higher education: An ensemble learning approach with feature importance analysis. *Journal of Information and Technology, 5*(4), 31–40. https://doi.org/10.70619/vol5iss4pp31-40

Osemwegie, E. E., & Amadin, F. I. (2023). Student dropout prediction using machine learning. *FUDMA Journal of Sciences, 7*(6), 347–353. https://doi.org/10.33003/fjs-2023-0706-2103

Ouyang, F., Wu, M., Zheng, L., Zhang, L., & Jiao, P. (2023). Integration of artificial intelligence performance prediction and learning analytics to improve student learning in online engineering course. *International Journal of Educational Technology in Higher Education, 20*(1), Article 4. https://doi.org/10.1186/s41239-022-00372-4

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, Article n71. https://doi.org/10.1136/bmj.n71

Patacsil, F. F. (2020). Survival analysis approach for early prediction of student dropout using

Revista Scientific - Artículo de Revisión Sistemática - Arbitrado - Volumen 11, Número 40 - Mayo-Julio 2026 - pp. 93-111
pp. BA2016000002 - e-ISSN: 2542-2987 - ISNI: 0000 0004 6045 0361

DOI: https://doi.org/10.29394/Scientific.issn.2542-2987.2026.11.40.6.93-111
OAI-PMH: http://www.indteca.com/ojs/index.php/Revista_Scientific/oai

enrollment student data and ensemble models. *Universal Journal of Educational Research, 8*(9), 4036–4047. https://doi.org/10.13189/ujer.2020.080929

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. JMLR.

Perkash, A., Shaheen, Q., Saleem, R., Rustam, F., Villar, M. G., Alvarado, E. S., … Ashraf, I. (2024). Feature optimization and machine learning for predicting students' academic performance in higher education institutions. *Education and Information Technologies, 29*(16), 21169–21193. https://doi.org/10.1007/s10639-024-12698-9

Polo, V. J. (2024). *Modelo predictivo basado en machine learning supervised y la deserción estudiantil en centros de educación superior tecnológicos públicos de la región La Libertad* [Tesis doctoral]. Universidad Nacional del Santa.

Putra, L. G. R., Prasetya, D. D., & Mayadi. (2025). Student dropout prediction using Random Forest and XGBoost method. *INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi, 9*(1), 147–157. https://doi.org/10.29407/intensif.v9i1.21191

Quintana, J. O. (2023). *Análisis predictivo de la deserción estudiantil en los alumnos de la Universidad Nacional de Moquegua entre 2009 y 2019, usando machine learning* [Tesis doctoral]. Universidad Nacional del Altiplano.

Ramírez, P. E., & Grandón, E. E. (2018). Predicción de la deserción académica en una universidad pública chilena a través de la clasificación basada en árboles de decisión con parámetros optimizados. *Formación Universitaria, 11*(3), 3–10. https://doi.org/10.4067/S0718-50062018000300003

Ramírez, R. C., & Carcausto-Calla, W. (2024). Factors associated with student dropout in Latin American universities: Scoping review. *Journal of Educational and Social Research, 14*(2), Article 62. https://doi.org/10.36941/jesr-2024-0026

Rodríguez, C. L., García, E., Brito, J., Durántez, F. Á., Silva, E., & Crespo, J. (2023). Forecasting of post-graduate students' late dropout based on the optimal probability threshold adjustment technique for imbalanced data. *International Journal of Emerging Technologies in Learning (iJET), 18*(04), 120–155. https://doi.org/10.3991/ijet.v18i04.34825

Romero, S., & Liao, X. (2025). Statistical and machine learning models for predicting university dropout and scholarship impact. *PLOS ONE, 20*(6), Article e0325047. https://doi.org/10.1371/journal.pone.0325047

Shafiq, D. A., Marjani, M., Habeeb, R. A. A., & Asirvatham, D. (2022). Student retention using educational data mining and predictive analytics: A systematic literature review. *IEEE Access, 10*, 72480–72503. https://doi.org/10.1109/ACCESS.2022.3188767

Sharma, R., Shrivastava, S. S., & Sharma, A. (2023). Predicting student performance using educational data mining and learning analytics technique. *Journal of Intelligent Systems and Internet of Things, 10*(2), 24–37. https://doi.org/10.54216/jisiot.100203

Shrestha, S., & Pokharel, M. (2020). Data mining applications used in education sector. *Journal of Education and Research, 10*(2), 27–51. https://doi.org/10.3126/jer.v10i2.32721

Shrestha, S., & Pokharel, M. (2021). Educational data mining in Moodle data. *International Journal of Informatics and Communication Technology (IJ-ICT), 10*(1), 9–18. https://doi.org/10.11591/ijict.v10i1.pp9-18

Sihare, S. R. (2024). Student dropout analysis in higher education and retention by artificial intelligence and machine learning. *SN Computer Science, 5*(2), Article 202. https://doi.org/10.1007/s42979-023-02458-w

Sulak, S. A., & Koklu, N. (2024). Predicting student dropout using machine learning algorithms. *Intelligent Methods in Engineering Sciences, 3*(3), 91–98. https://doi.org/10.58190/imiens.2024.103

Tao, H. (2024). Educational data mining for student performance prediction: Feature selection and model evaluation. *Journal of Electrical Systems, 20*(3), 1063–1074. https://doi.org/10.52783/jes.3434

Urbina-Nájera, A. B., Téllez-Velázquez, A., & Cruz, R. (2021). Patrones que identifican a estudiantes universitarios desertores aplicando minería de datos educativa. *Revista Electrónica de Investigación Educativa, 23*, 1–15. https://doi.org/10.24320/redie.2021.23.e29.3918

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), Article 2. https://doi.org/10.1007/s44163-023-00079-z

Winarsih, W., Sutanto, H., & Widodo, A. P. (2025). Implementation of the ensemble machine learning algorithm for student dropout prediction analysis. *Jurnal Sistem Informasi Bisnis, 15*(2), 159–166. https://doi.org/10.14710/vol15iss2pp159-166

Wolff, R. F., Moons, K. G. M., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S. (2019). PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine, 170*(1), 51–58. https://doi.org/10.7326/M18-1376

Xavier, M., & Meneses, J. (2021). The tensions between student dropout and flexibility in learning design: The voices of professors in open online higher education. *The International Review of Research in Open and Distributed Learning, 22*(4), 72–88. https://doi.org/10.19173/irrodl.v23i1.5652

Revista Scientific - Artículo de Revisión Sistemática - Arbitrado - Volumen 11, Número 40 - Mayo-Julio 2026 - pp. 93-111
pp. BA2016000002 - e-ISSN: 2542-2987 - ISNI: 0000 0004 6045 0361

DOI: https://doi.org/10.29394/Scientific.issn.2542-2987.2026.11.40.6.93-111
OAI-PMH: http://www.indteca.com/ojs/index.php/Revista_Scientific/oai

Xu, C., Zhu, G., Ye, J., & Shu, J. (2022). Educational data mining: Dropout prediction in XuetangX MOOCs. *Neural Processing Letters, 54*(4), 2885–2900. https://doi.org/10.1007/s11063-022-10745-5

Ye, C., & Biswas, G. (2014). Early prediction of student dropout and performance in MOOCs using higher granularity temporal information. *Journal of Learning Analytics, 1*(3), 169–172. https://doi.org/10.18608/jla.2014.13.14

Yükseltürk, E., Özekes, S., & Türel, Y. K. (2014). Predicting dropout student: An application of data mining methods in an online education program. *European Journal of Open, Distance and E-Learning, 17*(1), 118–133. https://doi.org/10.2478/eurodl-2014-0008

Zając, T., Perales, F., Tomaszewski, W., Xiang, N., & Zubrick, S. R. (2024). Student mental health and dropout from higher education: An analysis of Australian administrative data. *Higher Education, 87*(2), 325–343. https://doi.org/10.1007/s10734-023-01009-9

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

Reseña Biográfica de los autores

Jesús Eduardo Zambrano Loor | jezambrano@sangregorio.edu.ec

Es Ingeniero en Sistemas Informáticos por la Universidad Técnica de Manabí (UTM). Actualmente cursa la Maestría en Gestión y Analítica de Datos en la Universidad San Gregorio de Portoviejo (USGP), a la espera de la fecha de defensa de su trabajo de investigación. Se desempeña como Supervisor del Centro de Monitoreo del Departamento de Redes de Datos y Conectividad de la USGP (Manabí, Ecuador) desde 2024. Sus líneas de investigación se orientan a la analítica de datos aplicada a la educación superior.

Luis Daniel Muñoz Mendoza | ldmunoz@sangregorio.edu.ec

Es Ingeniero en Sistemas Informáticos por la Universidad Técnica de Manabí (UTM) y Magíster en Telecomunicaciones por la Universidad Católica de Santiago de Guayaquil (UCSG). Es docente de la Universidad San Gregorio de Portoviejo (USGP) desde 2015 y Jefe del Departamento de Redes de Datos y Conectividad de la misma institución desde 2019. Sus líneas de investigación se centran en redes de datos, conectividad institucional y tecnologías aplicadas a la gestión universitaria.

Jodamia Uridisnalda Murillo Rosado | jumurillo@sangregorio.edu.ec

Es Ingeniera en Sistemas Informáticos por la Universidad Técnica de Manabí (UTM) y Magíster en Informática Empresarial por la Universidad Regional Autónoma de los Andes (Uniandes). Actualmente cursa el doctorado en Ciencias Sociales y Jurídicas en la Universidad de Córdoba (UCO, España), a la espera de la fecha de defensa de su trabajo doctoral. Es profesora de la Universidad San Gregorio de Portoviejo (USGP) desde 2009 y miembro de su Departamento de Evaluación y Acreditación desde 2018. Sus líneas de investigación abarcan la evaluación institucional y la gestión de la calidad educativa en la educación superior.

Agradecimientos. Los autores agradecen a la Universidad San Gregorio de Portoviejo por el respaldo institucional brindado para el desarrollo del programa de Maestría en Gestión y Analítica de Datos, en cuyo marco se enmarca la

DOI: https://doi.org/10.29394/Scientific.issn.2542-2987.2026.11.40.6.93-111
OAI-PMH: http://www.indteca.com/ojs/index.php/Revista_Scientific/oai

presente investigación, así como al Departamento de Redes de Datos y Conectividad y al Departamento de Evaluación y Acreditación de dicha institución, por las facilidades otorgadas para la revisión y sistematización de la literatura científica utilizada en este estudio.

Declaraciones de los autores

Contribución de los autores (Taxonomía CRediT). Jesús Eduardo Zambrano Loor: Conceptualización, Investigación, Metodología, Curación de datos, Análisis formal, Validación, Escritura - borrador original y Escritura - revisión y edición. Luis Daniel Muñoz Mendoza: Conceptualización, Investigación, Metodología, Curación de datos, Supervisión, Validación y Escritura - revisión y edición. Jodamia Uridisnalda Murillo Rosado: Conceptualización, Investigación, Metodología, Curación de datos, Supervisión, Validación y Escritura - revisión y edición.

Financiamiento. Esta investigación no recibió financiamiento externo.

Conflicto de intereses. Los autores declaran no tener conflicto de intereses.

Declaración de disponibilidad de datos. El presente estudio se basó en la revisión de literatura científica publicada y de acceso público. No se generaron conjuntos de datos primarios adicionales a los ya citados en la lista de referencias de este artículo.

Declaración de uso de Inteligencia Artificial. Se utilizó una herramienta de inteligencia artificial (Claude, Anthropic) únicamente en la fase de revisión editorial del manuscrito, para apoyar la corrección de estilo y la verificación del cumplimiento del formato APA 7.ª edición; no se empleó en la búsqueda, el cribado, la selección, la extracción de datos ni la evaluación de calidad de los estudios. Todo el contenido fue revisado y verificado por los autores, quienes asumen la responsabilidad final sobre los datos, los resultados y las conclusiones del estudio.

Aprobación ética y consentimiento informado. Este estudio no requirió aprobación de un comité de ética, dado que consiste en una revisión sistemática de literatura científica ya publicada, sin intervención directa de seres humanos ni recolección de datos primarios.