UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Tecnura

https://revistas.udistrital.edu.co/ojs/index.php/Tecnura/issue/view/1254

DOI: https://doi.org/10.14483/22487638.23670

REVISIÓN

Modelos de inteligencia artificial en minería de datos educativos para predecir la deserción en Educación Superior: una revisión integral

Artificial Intelligence Models in Educational Data Mining for Predicting Dropout in Higher Education: A Comprehensive Review

José Leonardo Pérez Niño [1], Oscar Eduardo Gualdrón Guerrero [2] y Diego José Barrera Oliveros [3]

Fecha de Recepción: 15 de agosto de 2024 **Fecha de Aceptación**: 15 de diciembre de 2024

Cómo citar: Pérez-Niño, J., Gualdrón-Guerrero, O., y Barrera-Oliveros, D. (2024). Modelos de inteligencia artificial en minería de datos educativos para predecir la deserción en educación superior: una revisión integral. *Tecnura*, 28(82), 134-155. https://doi.org/10.14483/22487638.23670

Resumen

Contexto: la deserción estudiantil en educación superior afecta la calidad y permanencia educativa. La inteligencia artificial (IA) emerge como herramienta clave para predecir y prevenir este fenómeno, a través de modelos aplicados en minería de datos educativos.

Objetivo: analizar modelos de IA utilizados para predecir la deserción en educación superior, identificar variables frecuentes y evaluar la precisión y exactitud de los algoritmos aplicados.

Metodología: se realizó una revisión integral de literatura científica, centrada en estudios recientes que aplican IA en entornos educativos. Se compararon modelos según su rendimiento y frecuencia de uso, así como las variables más empleadas en los procesos predictivos.

Resultados: los modelos más destacados por su rendimiento fueron bosques aleatorios, redes neuronales artificiales y redes profundas. Las variables académicas, demográficas y socioeconómicas fueron las más significativas en los modelos predictivos.

Conclusiones: la IA permite anticipar riesgos de deserción con alta eficacia. Su aplicación depende de factores técnicos y contextuales. Se recomienda profundizar en nuevas variables y combinar enfoques tradicionales con aprendizaje profundo para mejorar la capacidad predictiva.

[1]Ingeniero Mecatrónico, Magister en controles industriales (c), Facultad de Ingenierías y arquitectura, Universidad de Pamplona, Pamplona, Colombia. Correo electrónico: jose.perez5@unipamplona.edu.co

[2]Ingeniero Electrónico, PhD en Ingeniería Electrónica. Profesor titular, Facultad de Ingenierías y arquitectura, Universidad de Pamplona, Pamplona, Colombia. Correo electrónico: oscar.gualdron@unipamplona.edu.co

[3]Ingeniero Mecatrónico, Mg en controles Industriales. Doctor en automática (c). Profesor de tiempo completo, Facultad de Ingenierías y arquitectura, Universidad de Pamplona, Pamplona, Colombia. Correo electrónico: diego.barrera@unipamplona.edu.co

Financiamiento: Este estudio hace parte de un proyecto de investigación aprobado por la Universidad de Pamplona, con financiación de recursos propios y apoyo institucional.

Palabras clave: Deserción estudiantil, algoritmos de aprendizaje automático, modelos predictivos, inteligencia artificial

Abstract

Context: Student dropout in higher education affects both educational quality and student retention. Artificial intelligence (AI) emerges as a key tool to predict and prevent this phenomenon through models applied in educational data mining.
Objective: To analyze AI models used to predict dropout in higher education, identify frequently used variables, and evaluate the precision and accuracy of the applied algorithms.
Methodology: A comprehensive review of scientific literature was conducted, focusing on recent studies that apply AI in educational settings. Models were compared based on their performance and frequency of use, as well as the most commonly employed variables in predictive processes.
Results: The models that stood out for their performance were Random Forests (RF), Artificial Neural Networks (ANN), and Deep Neural Networks (DNN). Academic, demographic, and socioeconomic variables proved to be the most significant in predictive models.
Conclusions: AI enables the anticipation of dropout risks with high effectiveness. Its application depends on technical and contextual factors. It is recommended to explore new variables and combine traditional approaches with deep learning to enhance predictive capacity.
Funding: This study is part of a research project approved by the University of Pamplona, funded through personal resources and institutional support.

Keywords: Student dropout, machine learning algorithms, predictive models, artificial intelligence.

Introducción

La deserción estudiantil en educación superior representa un desafío global con impactos significativos en el desarrollo académico, económico y social. Diferentes estudios han identificado múltiples factores que inciden en la permanencia de los estudiantes, tales como el rendimiento académico, las condiciones socioeconómicas y la motivación personal. En este contexto, la aplicación de técnicas de IA ha surgido como una alternativa eficiente para predecir y mitigar este fenómeno.

En este artículo se presenta una revisión integral sobre la implementación de modelos de IA en la predicción del riesgo de deserción estudiantil. Se analizan diversas metodologías utilizadas en estudios recientes, destacando los enfoques de aprendizaje automático más efectivos. Asimismo, se examinan las variables más influyentes en la predicción, los desafíos técnicos en la implementación de estos modelos y recomendaciones para próximos trabajos.

Predicción de la deserción estudiantil

Los sistemas de alerta temprana, diseñados con modelos de aprendizaje automático (MAA) para identificar a los estudiantes en riesgo de deserción, pueden mejorar los mecanismos de focalización y conducir a intervenciones de política social eficientes en la educación (Colak Oz *et al.*, 2023), en donde el uso de la IA ofrece una educación personalizada, según las características de cada estudiante (González Castro *et al.*, 2019).

Se plantea que una de las principales causas que llevan a la deserción es el rendimiento académico de los estudiantes, el cual se reconoce como uno de los principales problemas que enfrentan las instituciones de Educación Superior, debido a la alta tasa de fracaso en algunas materias (Hoyos y Caicedo Castro, 2024). Por esta razón, es objeto de investigación en un gran número de los artículos revisados, cuyos objetivos consisten en predecir el rendimiento académico, mejorar las calificaciones a futuro y disminuir los índices de deserción.

Deserción estudiantil

La deserción es quizás uno de los fenómenos que más está afectando los sistemas de educación (Díaz, 2009). Se asume la deserción como el estado de un estudiante que, de manera voluntaria o forzosa, no registra matrícula por dos o más periodos académicos consecutivos del programa en el que se matriculó; y no se encuentra como graduado, o se ha retirado por motivos disciplinarios (Gutiérrez *et al.*, 2021). En 2017, tras un boletín del Observatorio de Educación Superior de Medellín (2017) se ratificó que en Latinoamérica la deserción está entre el 40 % y el 75 %.

En Colombia, el encargado de estimar la deserción universitaria es el Sistema para la Prevención de la Deserción de la Educación Superior (SPADIES) (Ministerio de Educación Nacional [MEN], 2025), el cual en 2024 detalló en su informe que para el último periodo de estudio 2024-2, se registró una tasa de deserción de 12,3 %. En 2014, la Universidad de los Andes detalló que los motivos por los cuales los estudiantes desertaban eran principalmente socioeconómicos, académicos e institucionales (tabla 1).

Sin embargo, tras un estudio se concluyó que los factores más comunes y ampliamente utilizados para predecir el rendimiento de los estudiantes en la Educación Superior son sus calificaciones previas y su rendimiento en clase, su actividad de e-learning, su demografía y su información social (Abu Saa *et al.*, 2019).

Tabla 1. Motivos de deserción

Individuales	Socioeconómicos	Académicos	Institucionales
- Edad, género y estado civil	- Estrato social	- Orientación profesional	- Normatividad académica
- Calamidad y/o problema doméstico	- Situación laboral del estudiante	- Tipo de colegio de secundaria	- Becas y formas de financiamiento
- Integración social	- Situación laboral de los padres	- Rendimiento académico superior	- Recursos universitarios
- Expectativas no satisfechas	- Dependencia económica	- Métodos de estudio	- Relaciones con el profesorado y con demás estudiantes
- Incompatibilidad horaria con actividades extraacadémicas	- Personas a cargo	- Calificación en el examen de admisión	- Grado de compromiso con la institución educativa
	- Nivel educativo de los padres	- Insatisfacción con el programa académico	- Calidad del programa
	- Entorno familiar	- Carga académica (número de materias al semestre)	
	- Entorno macroeconómico del país	- Repitencia	

Fuente: Universidad de los Andes, Facultad de Economía, CEDE (2014)

Minería de datos educativos

La minería de datos educativos (MDE) es el proceso de extraer información útil y patrones de una gran base de datos educativa (Mengash, 2020). Su objetivo es identificar los entornos que favorecen el aprendizaje para mejorar los resultados y obtener información sobre los fenómenos educativos (Guo *et al.*, 2015); todos estos datos pueden usarse para predecir la probabilidad de deserción, el rendimiento de los estudiantes e, inclusive, la admisión en las universidades.

Predecir el comportamiento y los resultados de aprendizaje del alumnado se considera una de las tareas más importantes del campo de MDE. El interés principal se centra en tres tipos de problemas predictivos (Tsiakmaki *et al.*, 2019):

- Predecir si los estudiantes aprobarán o no un curso.
- Identificar a los estudiantes que tienen mayor riesgo de abandonar un curso, un problema crucial en el aprendizaje a distancia.
- Estimar las calificaciones de los estudiantes en pruebas, exámenes o cursos específicos.

Esto puede dar como resultado que disminuyan los índices de deserción, mejorar rendimiento académico de los estudiantes y mejorar la toma de decisiones en las diferentes instituciones de educación para avanzar con el proceso de admisión de sus aspirantes.

Metodología

Este estudio se enmarca en una investigación de tipo documental y aplicada, ya que se fundamenta en el análisis de literatura científica y tiene como propósito identificar herramientas predictivas que puedan ser utilizadas por instituciones de educación superior para mitigar la deserción estudiantil. El enfoque metodológico corresponde a una revisión integral de literatura científica, centrada en estudios publicados entre 2015 y 2024, priorizando aquellas que presentan métricas de precisión y exactitud en la predicción de la deserción estudiantil. La selección de fuentes se basó en criterios de relevancia, impacto académico y aplicabilidad de los modelos. Asimismo, se identificaron las variables más utilizadas, los algoritmos más frecuentes y los marcos normativos asociados, con el fin de establecer un estado del arte riguroso y actualizado.

A continuación, se analizan las diferentes investigaciones relevantes realizadas en los últimos años sobre la minería de datos educativos, identificando la precisión y exactitud de los modelos con mayor frecuencia de aparición en los trabajos de investigación consultados, y de esta manera lograr identificar el modelo que tiene mejores resultados y mayores citaciones.

Dentro de los artículos se evalúa la precisión y exactitud que tienen los modelos, las cuales se utiliza con el fin de conocer las predicciones realizadas de manera correcta, son medidas de probabilidad de que un documento clasificado o predicho en la clase que corresponda realmente sea esa (Paniagua Medina *et al.*, 2023), con la distinción que la precisión es el grado de concordancia entre los resultados de mediciones repetidas de un mismo objeto y la exactitud hace referencia al grado de concordancia entre una medición y el valor real de una cantidad medida.

Árbol de decisión (AD)

Un árbol de decisión se estructura a partir de un nodo raíz, que se extiende a través de nodos intermedios, conocidos como nodos hoja, hasta llegar a un nodo final (Khan *et al.*, 2019). Esta técnica se emplea para identificar patrones y reglas, dividiendo y segmentando de manera sistemática la información contenida en la base de datos (Agudelo, 2023).

Tabla 2. Precisión y exactitud AD

Autor y referencia	Precisión (%)	Exactitud (%)
Avila Pérez	58.3	58.33
Ramírez P.	87.27	-
Melo A.	88.89	86.81
Vives L.	100	100
Mendes R.	96.2	85.6
Khairy D.	98.7	99.0
Khan I.	-	88.0
Kostopoulos G.	-	85.19
Caicedo-Castro	66.54	66.54
Abideen Zain	94	97
Nasa P.	98.24	95.78
Haron N.	95.8	95.50
Agudelo O.	97.63	93.81
Li X.	80	95

Fuente: elaboración propia

Bosques aleatorios (BA)

Están compuestos por múltiples árboles de decisión generados de manera aleatoria. Se caracterizan por su alta precisión, no requieren escalado ni codificación de variables categóricas y requieren un ajuste mínimo de parámetros (Abideen *et al.*, 2023).

Tabla 3. Precisión y exactitud BA

Autor y referencia	Precisión %	Exactitud %
Mendes	97,6	88.9
Vives L.	100	100
Nasa P.	98.25	97.73
Caicedo-Castro	71.67	73.91

Abideen Zain	95	93
Agudelo O.	93.86	93.0
Khairy D.	98.7	99.0
Li X.	97	98
Hien	-	83.21

Fuente: elaboración propia

Clasificador bayesiano ingenuo (BI)

Los métodos del Clasificador bayesiano ingenuo o Naive Bayes son un conjunto de algoritmos de aprendizaje supervisado que se basan en la aplicación del teorema de Bayes con la suposición "Naive"de independencia condicional entre cada par de características dado el valor de la variable de clase (Paniagua Medina *et al.*, 2023).

Tabla 4. Precisión y exactitud BI

Autor y referencia	**Precisión (%)**	**Exactitud (%)**
Ávila	57.0	57.83
Guo B.	-	20.7
Khan I.	-	84.0
Abideen Zain	49	63
Haron N.	-	94.38
Khairy D.	94.0	94.0

Fuente: elaboración propia

Red neuronal profunda (RNP)

Las redes neuronales son modelos utilizados para clasificación, regresión y agrupamiento, diseñados a partir de la inspiración en el funcionamiento del cerebro humano. Sus nodos están interconectados, permitiendo el intercambio de información de manera similar a cómo las neuronas biológicas se comunican mediante dendritas y axones (Aslam *et al.*, 2021).

La RNP destaca por su capacidad para modelar relaciones no lineales complejas mediante la incorporación de múltiples capas ocultas. Su propósito es permitir que la red aprenda automáticamente características relevantes para tareas de clasificación o regresión (Li *et al.*, 2024). Las redes neuronales profundas pueden modelar mejor el comportamiento de los estudiantes y obtienen un buen rendimiento de predicción en comparación con los algoritmos de aprendizaje automático tradicionales.

Tabla 5. Precisión y exactitud RNP

Autor y referencia	Precisión (%)	Exactitud (%)
Vives L.	96.8	97.7
Aslam	94.9	93
Hussain	96	95.34
Li X.	94	94
Liu T.	90.6	92.5
Yang & Bai	90.7	86.6
Guo B.	77.2	77.2

Fuente: elaboración propia

Red neuronal convolucional (RNC)

Son ampliamente empleadas en tareas de predicción, ya que ofrecen diversas configuraciones que deben ajustarse y optimizarse para lograr un rendimiento preciso. Factores como la función de activación, la cantidad de capas, el número de neuronas por capa y la elección de hiperparámetros juegan un papel crucial en la eficacia del modelo (Akour *et al.*, 2020).

La RNC procesa imágenes mediante convoluciones, permitiendo calcular respuestas basadas en la relación entre píxeles cercanos y utilizando como entrada una imagen representada en una matriz. Por su potencial en el análisis de datos visuales, RNC es un modelo óptimo para predecir el rendimiento académico de los estudiantes, categorizándolos en aprobado o reprobado (Poudyal *et al.*, 2022). Este enfoque representa una innovación en la MDE debido a la forma en que se implementan las arquitecturas RNC en el análisis.

Tabla 6. Precisión y exactitud RNC

Autor y referencia	Precisión (%)	Exactitud (%)
Akour	95.6	95.5
Kavipriya	91.25	92.41
Poudyal	-	88

Fuente: elaboración propia

Redes de memoria a corto plazo (RMCP)

Estas procesan y almacenan información de entrada mediante tres unidades de compuerta, permitiendo que el modelo retenga únicamente los datos más relevantes y, de este modo,

optimice el uso de la memoria. Gracias a esta capacidad, RMCP es eficaz en la resolución de problemas de dependencia a largo plazo. Además, su aplicación es útil para identificar patrones temporales en series de datos no lineales (Chen *et al.*, 2023).

Tabla 7. Precisión y exactitud RNC-RMCP

Autor y referencia	Precisión (%)	Exactitud (%)
Vives L.	97	97.7
Aljohani	93.46	95.23
Xie	92.07	89.12
Hien	-	86.26
Chen	84	85

Fuente: elaboración propia

K vecinos más cercanos (K-VMC)

El método de clasificación es una de las técnicas más utilizadas en el aprendizaje automático. Su principio fundamental es clasificar nuevos datos identificando las instancias más similares dentro del conjunto de entrenamiento y basando la predicción en sus etiquetas. Debido a su simplicidad y facilidad de implementación, *k*-VMC ha sido ampliamente aplicado en diversos ámbitos, como sistemas de recomendación, búsqueda semántica e identificación de anomalías (Abideen *et al.*, 2023).

Tabla 8. Precisión y exactitud k-VMC

Autor y referencia	Precisión (%)	Exactitud (%)
Vives L.	97	98.1
Abideen Zain	80	84
Haron N.	-	88.76
Khairy D.	90.0	89.6

Fuente: elaboración propia

Regresión logística (RL)

Emplea una función logística para modelar la relación entre una combinación lineal de las variables de entrada y sus respectivos pesos. El modelo ajusta un clasificador optimizando la función objetivo, basada en la verosimilitud logarítmica de los datos de entrenamiento, predice un resultado binario (Caicedo Castro, 2023).

Tabla 9. Precisión y exactitud de RL

Autor y referencia	Precisión (%)	Exactitud (%)
Vives L.	99.6	99.3
Caicedo-Castro	84.16	75.73
Mendes	96.7	89.1
Li X.	43	38
Xie	81.06	88.47

Fuente: elaboración propia

Redes neuronales artificiales (RNA)

Se compone de tres capas principales: la capa de entrada, la capa oculta y la capa de salida. Para su diseño, es fundamental considerar tres aspectos clave: la cantidad de capas ocultas, el número de neuronas en cada una y la función de activación utilizada en las neuronas (Rabelo y Zárate, 2024).

Tabla 10. Precisión y exactitud de RNA

Autor y referencia	Precisión (%)	Exactitud (%)
Khairy D.	96.0	96.4
Mendes	97.1	87.2
Xie	94.67	94.61

Fuente: elaboración propia

Máquinas de soporte vectorial (MSV)

Es un MAA supervisado que utiliza algoritmos de clasificación para abordar problemas de categorización binaria. Este enfoque es especialmente eficaz para trabajar con datos dispersos, destacándose por su simplicidad y fiabilidad en la clasificación (Abideen *et al.*, 2023).

Tabla 11. Precisión y exactitud de MSV

Autor y referencia	Precisión (%)	Exactitud (%)
Burgos C.	36.73	62.50
Guo	-	48.4
Caicedo-Castro	85.17	77.64

Abideen Zain	84	59
Nasa P.	91.83	90.97

Fuente: elaboración propia

Resultados

A manera de resumen, a continuación, se incluye el promedio de todas las tablas previas, demostrando que los BA tienen buenos resultados en la relación de precisión, exactitud y frecuencia de aparición. Por su parte las RNA tienen los mejores resultados en cuanto precisión y exactitud, sin embargo, son pocos citados en los trabajos consultados. Cabe destacar que el AD fue el modelo con mayores citaciones, sin embargo, sus resultados de precisión y promedio no son los mejores.

Tabla 12. Precisión promedio de todos los modelos

Modelo	Precisión Promedio (%)	Exactitud Promedio (%)	Frecuencia de aparición
Árbol de decisión	88.46	88.2 %	14
Bosques Aleatorios	94.01	91.86	9
Red neuronal profunda	91.46	90.90	7
Clasificador bayesiano ingenuo	66.67	68.99	6
Redes de memoria a corto plazo	91.63	90.662	5
Regresión logística	80.904	78.12	5
Máquinas de soporte vectorial	74.43	67.702	5
K Vecinos más cercanos	89	90.115	4
Red neuronal convolucional	93.43	91.97	3

Redes neuronales artificiales	95.92	92.73	3

Fuente: elaboración propia

Identificación de variables

Dentro de los trabajos de investigación consultados, gran parte de las razones por las cuales los modelos obtienen buenos resultados se debe a la calidad de las variables que se utilizan, por ello es imperativo que el presente artículo tenga como riqueza añadida la presentación de la figura **1**, en la cual se exponen las variables con mayores frecuencias de aparición, además de dar claridad de los autores que abordan su problema de investigación con cada variable en particular.

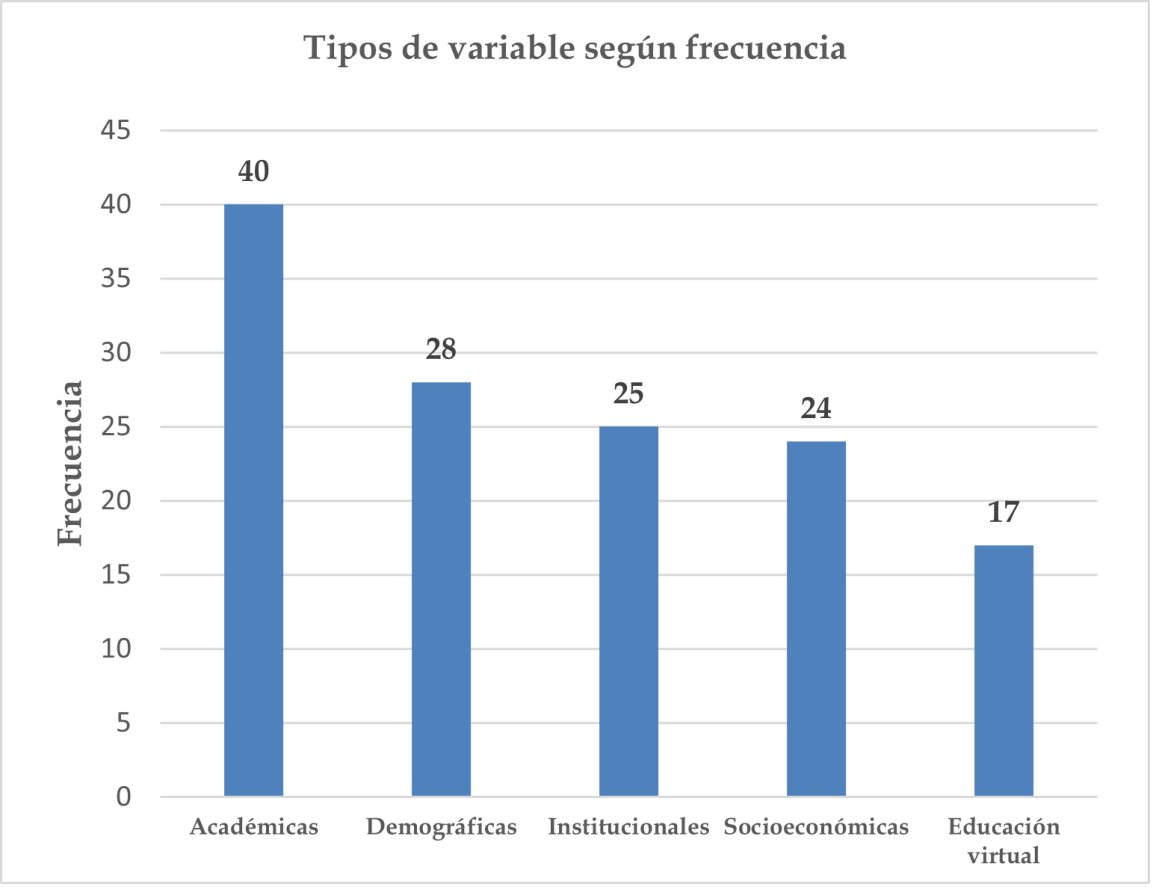

Figura 1. Tipos de variable según frecuencia de aparición en los artículos

Localización del área de estudio

Las variables con mayores citaciones son las académicas, dentro de ellas encontramos 28 autores que toman en cuenta el historial académico del estudiante y las calificaciones que tienen al momento del estudio para entrenar el modelo; algunos de ellos son los trabajos de Guo *et al*. (2015); Khan *et al*. (2019); Agudelo (2023); Aslam *et al*. (2021).

Por otra parte, algunos autores mencionan la importancia de reconocer el promedio académico en la Educación Media como variable predictiva, en paralelo, en Colombia, las Pruebas Saber 11 adelantadas por el Instituto Colombiano para la Evaluación de la Educación (Icfes), constituye un elemento de análisis (Hoyos y Caicedo Castro, 2024; Guo *et al*., 2015; Khan *et al*., 2019; Agudelo, 2023; Caicedo Castro, 2023; Ramírez Gualdrón, 2018; Haron *et al*., 2024; Hien *et al*., 2020; Iqbal *et al*., 2019; Roslan *et al*., 2024).

En algunos estudios existen patrones asociados al buen o mal desempeño académico de los estudiantes, con respecto al puntaje global obtenido en las competencias genéricas del Examen Saber Pro (Timarán Buchely y Timarán Pereira, 2023).

Por último, en el apartado de variables académicas, Agudelo (2023) y Ávila Pérez (2021) analizan el número de asignaturas reprobadas del estudiante para saber si desertará o no.

Variables demográficas

Las más consultadas son edad y género, y son objeto de estudio por los siguientes autores: Aljohani *et al*. (2019); Akour *et al*. (2020); Aslam *et al*. (2021); Guo *et al*. (2015); Hoyos y Caicedo Castro (2024); Paniagua *et al*. (2019); Poudyal *et al*. (2022); Roslan *et al*. (2024); Xie (2021); Zhang *et al*. (2010).

Por su parte, el lugar de procedencia del estudiante, así como la vivienda en la que reside, se contempla por seis estudios como un elemento importante para el entrenamiento del modelo: Aljohani *et al*. (2019); Aslam *et al*. (2021); Hoyos y Caicedo Castro (2024); Roslan *et al*. (2024); Xie (2021); Zhang *et al*. (2010).

Factores institucionales y personales

La asistencia del estudiante a clases es analizada en seis documentos, en los cuales se recalca la importancia de este factor en la permanencia estudiantil: Agudelo (2023); Akour *et al*. (2020); Albreiki *et al*. (2021); Ali Khan *et al*. (2024); Kostopoulos *et al*. (2019); Rabelo y Zárate (2024).

En otros trabajos se menciona que los estudiantes con mayor participación en actividades académicas, ya sean extracurriculares, de liderazgo, entre otros, tienen mayores posibilidades de no desertar (Agudelo, 2023; Akour *et al.*, 2020; Albreiki *et al.*, 2021; Aslam *et al.*, 2021; Haron *et al.*, 2024; Hoyos y Caicedo Castro, 2024; Nurhana Roslan *et al.*, 2024).

Agudelo (2023); Akour *et al.* (2020), y Rabelo y Zárate (2024) mencionan que la satisfacción de los estudiantes con su proceso académico en la universidad en la que están matriculados tiene relevancia en el estudio. Para Ávila Pérez (2021); Haron *et al.* (2024); Hien *et al.* (2020), y Mariano *et al.* (2022), las relaciones interpersonales entre el estudiante y los docentes o los compañeros de clases tienden a ser una causa de deserción.

Finalmente, Guo *et al.* (2015) y Haron *et al.* (2024) abordan la motivación y la personalidad del estudiante como dos características importantes en el análisis del estudio y posterior creación del modelo.

Variables socioeconómicas

Son importantes para la predicción de la deserción. En nueve documentos se afirma que la situación económica del estudiante es trascendental para reconocer si este continuará con sus estudios: Hoyos y Caicedo Castro (2024); Agudelo (2023); Alalawi *et al.* (2023); Albreiki *et al.* (2021); Aslam *et al.* (2021); Ávila Pérez (2021); Rabelo y Zárate (2024); Ramírez y Grandón (2018), y Roslan *et al.* (2024).

También se menciona que el estado familiar y el nivel educativo de los padres influyen en la predicción, por ello, los siguientes autores lo incluyen en su investigación: Akour *et al.* (2020), Alalawi *et al.* (2023), Ávila Pérez (2021), Guo *et al.* (2015), Hoyos y Caicedo Castro (2024), Kavipriya y Sengaliappan (2021), Ramírez y Grandón (2018), Rabelo y Zárate (2024), Zhang *et al.* (2010). Por último, según Agudelo (2023), Rabelo y Zárate (2024), y Roslan *et al.* (2024), se deben tener en cuenta las becas recibidas por el estudiante para este pronóstico.

Comportamiento en educación virtual

En muchos estudios se analizan específicamente los cursos masivos abiertos en línea (CMAL). Para la permanencia en estos se consideran diferentes variables que permiten evaluar la usabilidad de un software a través de test con usuarios, en los cuales se monitoriza y observa el comportamiento de un grupo de usuarios, mientras estos desarrollan una serie de tareas en un espacio controlado (Chanchí Golondrino *et al.*, 2024), como el tiempo de conexión o número de clics, variables que son abordadas en los siguientes documentos: Alalawi *et al.* (2023); Aljohani *et al.* (2019); Alnasyan *et al.* (2024); Chen (2022); Kostopoulos *et al.* (2019); Li *et al.* (2022); Liu *et*

al. (2022); Poudyal *et al*. (2022), y Xie (2021).

También es importante mencionar que la participación en foros, las bases de datos consultadas por el estudiante y la participación general en los procesos en línea son tema de estudio por ocho escritos: Akour *et al*. (2020); Alnasyan *et al*. (2024); Hoyos y Caicedo Castro (2024); Li *et al*. (2022); Liu *et al*. (2022), y Zhang *et al*. (2010).

Otras variables

Teniendo en cuenta otras características como el factor innovador, es pertinente mencionar las siguientes variables:

- Escuela de procedencia (Ramírez y Grandón, 2018): se trata del lugar en donde el estudiante se graduó de Educación Media.
- Número de créditos matriculados (Poudyal *et al*., 2022): hace referencia a la carga académica por semestre.
- Programa al que está inscrito el estudiante (Agudelo, 2023): el artículo menciona que los diferentes programas tienen diferentes tipos de dificultad y ello afecta las probabilidades de deserción del estudiante.
- Estado civil (Agudelo, 2023).
- Discapacidad (Haron *et al*., 2024): se expresa el tipo de discapacidad del estudiante.
- Primera opción (Agudelo, 2023): en el artículo se pregunta si el programa que el estudiante está estudiando fue elegido como primera opción al matricularse a la universidad.
- Estado de salud (Aslam *et al*., 2021).
- Tiempo en iniciar la vida universitaria (Hien *et al*., 2020): cuánto tiempo tardó en ingresar a la Educación Superior desde que se graduó de la Educación Media.
- Habilidades blandas de cada estudiante (Kavipriya y Sengaliappan, 2021).

Futuras recomendaciones

Dentro de la revisión integral se plantean nuevos modelos alternativos con buenos resultados, como el uso del aprendizaje profundo con las redes neuronales artificiales, profundas y convolucionales. El uso de modelos que históricamente han sido utilizados para procesamiento de imágenes o audios han sido de utilidad para los procesos de predicción de la deserción generando nuevos enfoques en estos modelos, por lo cual se recomienda seguir indagando en

estos procesos de inteligencia artificial que pueden desarrollar cada vez procesos más eficientes e innovadores.

También se recomienda indagar un poco más en nuevas variables que sean determinantes dentro del proceso de predicción, profundizar y utilizar las otras características explicitas dentro del presente artículo; como evaluar las habilidades blandas, si la carrera elegida responde al deseo vocacional de la persona, entre otros, generando una recolección de datos más robusta y eficaz, tal cual lo expresa L. F. Vargas Tamayo *et al* (2013) donde se plantea construir una solución efectiva al problema de la información académica estudiantil ya que en el momento en que se implante y aplique una recolección de datos eficiente, se logrará hacer seguimiento y mejoras posteriores.

Conclusiones

El presente estudio ha evidenciado la relevancia de la inteligencia artificial en la predicción del riesgo de deserción estudiantil, destacando cómo distintas técnicas de aprendizaje automático han sido aplicadas con diversos niveles de éxito. A partir del análisis de los modelos presentados, se puede concluir que no existe un enfoque único y universalmente óptimo, sino que la efectividad de cada método depende de su precisión, exactitud y su aplicabilidad en diferentes contextos educativos.

Los BA emergen como una de las técnicas más equilibradas, con una precisión promedio del 94.01 % y una exactitud del 91.86 %, lo que indica su capacidad para ofrecer predicciones confiables y generalizables. Su alta precisión, combinada con su frecuencia de aparición en los estudios revisados (9 menciones), sugiere que es un modelo preferido para la identificación del riesgo de deserción debido a su estabilidad y facilidad de implementación. Por otro lado, las RNA registran la mayor precisión (95.92 %) y exactitud (92.73 %), pero su menor frecuencia de uso (3 menciones) sugiere que su implementación es más compleja y menos común en la literatura revisada. Esto indica que, aunque son altamente efectivas, su demanda computacional y la necesidad de grandes volúmenes de datos de entrenamiento pueden limitar su aplicabilidad en entornos donde los recursos sean limitados. Las RNP y las RMCP también han demostrado ser enfoques prometedores, con precisiones del 91.46 % y 91.63 %, respectivamente, lo que confirma el potencial del aprendizaje profundo en este ámbito. Sin embargo, su implementación sigue siendo menos frecuente en comparación con modelos más tradicionales como los AD, los cuales, a pesar de ser los más citados en la literatura (14 menciones), presentan un desempeño inferior con una precisión promedio del 88.46 %. Esto sugiere que su popularidad se debe más a su facilidad de interpretación que a su desempeño en la tarea de predicción.

El análisis de las variables utilizadas en los modelos muestra que las calificaciones e historial académico son los predictores más utilizados (28 menciones), seguidos por las características demográficas (20 menciones) y el estado familiar (12 menciones). Esto confirma que el rendimiento académico sigue siendo el factor más determinante en la deserción, aunque variables socioeconómicas y motivacionales podrían aportar valor si se integraran más en los modelos predictivos.

A pesar de estos avances, la implementación de estas técnicas aún enfrenta desafíos, como la necesidad de mejorar la calidad y diversidad de los datos, la interpretabilidad de los modelos más complejos, la adaptación de los algoritmos a diferentes contextos educativos y las voluntades de cada institución de educación superior a implementarlas dentro de sus procesos. Futuras investigaciones deberían centrarse en combinar enfoques tradicionales con técnicas de aprendizaje profundo, así como en la incorporación de variables menos exploradas, como la motivación y el comportamiento en entornos virtuales, para mejorar la capacidad predictiva de los modelos.

En conclusión, la inteligencia artificial ofrece soluciones innovadoras y altamente efectivas para la predicción de la deserción estudiantil, tal como se expresa en (Llanos Mosquera *et al.*, 2021) la implementación de la inteligencia artificial se ha convertido en una prioridad para la educación virtual, potenciando la forma de entender y comprender las necesidades específicas del estudiante., pero su implementación debe considerar factores técnicos, computacionales, de voluntades y éticos. La combinación de modelos robustos, el uso de datos más representativos y el desarrollo de estrategias adaptadas a cada contexto educativo permitirán avanzar hacia sistemas de predicción más precisos y aplicables, promoviendo la permanencia y éxito de los estudiantes en la educación superior.

Financiamiento

Este trabajo es resultado del proyecto de investigación de maestría titulado “Algoritmo basado en inteligencia artificial para la predicción del riesgo en la deserción estudiantil de la Universidad de Pamplona”, avalado por el programa de Maestría en Controles Industriales y por la Vicerrectoría de Investigaciones de la Universidad de Pamplona, en el marco del proceso de trabajo de grado.

Agradecimientos

Agradecemos a la Universidad de Pamplona por el respaldo institucional brindado para el desarrollo de esta investigación. Reconocemos la diligencia, compromiso y disposición de los

miembros de su comunidad académica, cuyo apoyo fue fundamental en las distintas etapas del trabajo. Su colaboración, tanto en el acceso a recursos como en el acompañamiento técnico y académico, fue clave para la consolidación de este estudio.

Asimismo, expresamos nuestro agradecimiento a la revista Tecnura de la Universidad Distrital Francisco José de Caldas por fomentar el ejercicio de la difusión del conocimiento científico, contribuyendo al fortalecimiento de la investigación en el ámbito académico nacional e internacional.

Referencias

Abideen, Z. U., Mazhar, T., Razzaq, A., Haq, I., Ullah, I., Alasmary, H., y Mohamed, H. G. (2023). Analysis of enrollment criteria in secondary schools using machine learning and data mining approach. *Electronics (Switzerland), 12*(3), 694. https://doi.org/10.3390/electronics12030694

Abu Saa, A., Al-Emran, M., y Shaalan, K. (2019). Factors affecting students' performance in higher education: a systematic review of predictive data mining techniques. *Technology, Knowledge and Learning, 24*(4), 567-598. https://doi.org/10.1007/s10758-019-09408-7

Agudelo, O. (2023). *Identificación de deserción temprana en la Universidad de Manizales con aprendizaje automático, como parte de la estrategia de prevención* [Tesis de maestría, Universidad de Manizales]. Repositorio Institucional UM. https://ridum.umanizales.edu.co/handle/20.500.12746/647

Akour, M., Sghaier, H. Al, y Al Qasem, O. (2020). The effectiveness of using deep learning algorithms in predicting students achievements. *Indonesian Journal of Electrical Engineering and Computer Science, 19*(1), 388. https://doi.org/10.11591/ijeecs.v19.i1.pp388-394

Alalawi, K., Athauda, R., y Chiong, R. (2023). Contextualizing the current state of research on the use of machine learning for student performance prediction: a systematic literature review. *Engineering Reports, 5*(12). https://doi.org/10.1002/eng2.12699

Albreiki, B., Zaki, N., y Alashwal, H. (2021). A systematic literature review of student' performance prediction using machine learning techniques. *Education Sciences, 11*(9), 552. https://doi.org/10.3390/educsci11090552

Ali Khan, M. A., Ma, H., Farhad, A., Mujeeb, A., Mirani, I. K., y Hamza, M. (2024). When LoRa meets distributed machine learning to optimize the network connectivity for green and intelligent transportation system. *Green Energy and Intelligent Transportation, 3*(3), 100204. https://doi.org/10.1016/j.geits.2024.100204

Aljohani, N. R., Fayoumi, A., y Hassan, S.-U. (2019). Predicting at-risk students using clickstream data in the virtual learning environment. *Sustainability, 11*(24), 7238. https://doi.org/10.3390/su11247238

Alnasyan, B., Basheri, M., y Alassafi, M. (2024). The power of deep learning techniques for predicting student performance in virtual learning environments: a systematic literature review. *Computers and Education: Artificial Intelligence, 6*, 100231. https://doi.org/10.1016/j.caeai.2024.100231

Aslam, N. M., Khan, I. U., Alamri, L. H., y Almuslim, R. S. (2021). An improved early student's academic performance prediction using deep learning. *International Journal of Emerging Technologies in Learning (IJET), 16*(12), 108-122. https://doi.org/10.3991/ijet.v16i12.20699

Ávila Pérez, M. L. (2021). *Modelo de predicción de deserción estudiantil, apoyado en tecnologías de data mining, en un curso de primera matrícula de la escuela ECBTI de la UNAD* [Tesis de maestría, Universidad Nacional Abierta y a Distancia]. Repositorio Institucional UNAD. https://repository.unad.edu.co/handle/10596/42544

Caicedo Castro, I. (2023). Course prophet: a system for predicting course failures with machine learning: a numerical methods case study. *Sustainability, 15*(18), 13950. https://doi.org/10.3390/su151813950

Chanchí Golondrino, G. E., Ospina Alarcón, M. A., y Campo Muñoz, W. Y. (2024). Construction of a virtual usability laboratory for conducting user tests during remote attendance. *Revista Colombiana de Tecnologías de Avanzada (RCTA), 2*(44), 35-44. https://doi.org/10.24054/rcta.v2i44.2713

Chen, G. (2022). Pinning control of complex dynamical networks. *IEEE Transactions on Consumer Electronics, 68*(4), 336-343. https://doi.org/10.1109/TCE.2022.3200488

Chen, H., Wang, Y., Li, Y., Lee, Y., Petri, A., y Cha, T. (2023). Computer science and non-computer science faculty members' perception on teaching data science via an experiential learning platform. *Education and Information Technologies, 28*(4), 4093-4108. https://doi.org/10.1007/s10639-022-11326-8

Colak Oz, H., Güven, Ç., y Nápoles, G. (2023). School dropout prediction and feature importance exploration in Malawi using household panel data: machine learning approach. *Journal of Computational Social Science, 6*(1), 245-287. https://doi.org/10.1007/s42001-022-00195-3

Díaz, A. (2009). Análisis sobre la deserción en la Educación Superior a distancia y virtual: el caso de la UNAD – Colombia. *Revista de Investigaciones UNAD, 8*(2), 117-149.

González Castro, Y., Peñaranda Peñaranda, M. M., y Manzano Durán, O. (2019). Innovaciones tecnológicas en las prácticas académicas virtuales. *Revista Colombiana de Tecnologías de Avanzada (RCTA)*, *1*(33), 69-77. https://doi.org/10.24054/rcta.v1i33.87

Guo, B., Zhang, R., Xu, G., Shi, C., y Yang, L. (2015). Predicting students performance in educational data mining. En *2015 International Symposium on Educational Technology (ISET)* (pp. 125-128). Wuhan, República Popular de China. https://doi.org/10.1109/ISET.2015.33

Gutiérrez A., D., Vélez Díaz, J. F., y López M., J. (2021). Indicadores de deserción universitaria y factores asociados. *EducaT: Educación Virtual, Innovación y Tecnologías*, *2*(1), 15-26. https://doi.org/10.22490/27452115.4738.

Haron, N. H., Mahmood, R., Amin, N. M., Ahmad, A., y Jantan, S. R. (2024). An Artificial Intelligence Approach to Monitor and Predict Student Academic Performance. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, *44*(1), 105–119. https://doi.org/10.37934/araset.44.1.105119

Hien, D. T. T., Thi, C., Kim, T., The, D., y Nguyen, C. (2020). Optimize the combination of categorical variable encoding and deep learning technique for the problem of prediction of vietnamese student academic performance. *International Journal of Advanced Computer Science and Applications*, *11*(11). https://doi.org/10.14569/IJACSA.2020.0111135

Hoyos, W., y Caicedo Castro, I. (2024). Tuning data mining models to predict secondary school academic performance. *Data*, *9*(7), 86. https://doi.org/10.3390/data9070086

Iqbal, Z., Qayyum, A., Latif, S., y Qadir, J. (2019). Early student grade prediction: an empirical study. En *2019 2nd International Conference on Advancements in Computational Sciences (ICACS)* (pp. 1-7). Lahore, Pakistán. https://doi.org/10.23919/ICACS.2019.8689136

Kavipriya, T., y Sengaliappan, M. (2021). Adaptive weight deep convolutional neural network (AWDCNN) classifier for predicting student's performance in job placement process. *Annals of the Romanian Society for Cell Biology*, *25*(1), 507-518.

Khan, I., Al Sadiri, A., Ahmad, A. R., y Jabeur, N. (2019). Tracking student performance in introductory programming by means of machine learning. En *2019 4th MEC International Conference on Big Data and Smart City (ICBDSC)* (pp. 1-6). Mascate, Omán. https://doi.org/10.1109/ICBDSC.2019.8645608

Kostopoulos, G., Karlos, S., y Kotsiantis, S. (2019). Multiview learning for early prognosis of academic performance: a case study. *IEEE Transactions on Learning Technologies*, *12*(2), 212-224. https://doi.org/10.1109/TLT.2019.2911581

L. F. Vargas Tamayo, H. R. Tocasuche González, & J. A. Tristancho Ortiz. (2013). Nuevo software para la administración y control académico de estudiantes en ingeniería industrial. *Tecnura*, *17*(1), 174–189. https://doi.org/10.14483/22487638.7247

Li, B., Lowell, V. L., Wang, C., y Li, X. (2024). A systematic review of the first year of publications on ChatGPT and language education: examining research on ChatGPT's use in language learning and teaching. *Computers and Education: Artificial Intelligence*, *7*, 100266. https://doi.org/10.1016/j.caeai.2024.100266

Liu, T., Wang, C., Chang, L., y Gu, T. (2022). Predicting *High-Risk Students Using Learning Behavior. Mathematics*, *10*(14), 2483. https://doi.org/10.3390/math10142483

Llanos Mosquera, J. M., Hidalgo Suarez, C. G., & Bucheli Guerrero, V. A. (2021). Una revisión sistemática sobre aula invertida y aprendizaje colaborativo apoyados en inteligencia artificial para el aprendizaje de programación. *Tecnura*, *25*(69), 196–214. https://doi.org/10.14483/22487638.16934

Mariano, A. M., Ferreira, A. B. de M. L., Santos, M. R., Castilho, M. L., y Bastos, A. C. F. L. C. (2022). Decision trees for predicting dropout in engineering course students in Brazil. *Procedia Computer Science*, *214*, 1113-1120. https://doi.org/10.1016/j.procs.2022.11.285

Mengash, H. A. (2020). Using data mining techniques to predict student performance to support decision making in university admission systems. *IEEE Access*, *8*, 55462-55470. https://doi.org/10.1109/ACCESS.2020.2981905

Ministerio de Educación Nacional. (12 de febrero de 2025). *Estadísticas de deserción y permanencia en Educación Superior (SPADIES 3.0). Indicadores 2021*. https://www.mineducacion.gov.co/sistemasinfo/spadies/secciones/Estadisticas-de-desercion/

Observatorio de Educación Superior de Medellín. (2017). *Deserción en la educación superior.* https://www.sapiencia.gov.co/wp-content/uploads/2017/11/5_JULIO_BOLETIN_ODES_DESERCION_EN_LA_EDUCACION_SUPERIOR.pdf

Paniagua Medina, J. J., Vargas Rodríguez, E., y Guzmán Cabrera, R. (2023). Aprendizaje automático y la Colección Reuters-21578 en la clasificación de documentos. *Revista Colombiana de Tecnologías de Avanzada (RCTA)*, *2*(40), 39-46. https://doi.org/10.24054/rcta.v2i40.2344

Poudyal, S., Mohammadi-Aragh, M. J., y Ball, J. E. (2022). Prediction of student academic performance using a hybrid 2D CNN model. *Electronics*, *11*(7), 1005. https://doi.org/10.3390/electronics11071005

Rabelo, A. M., y Zárate, L. E. (2024). 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

Ramírez, P. E., y 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

Roslan, N., Jamil, J. M., Mohd Shaharanee, I. N., y Sultan Alawi, S. J. (2024). Prediction of student dropout in malaysian's private higher education institute using data mining application. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, *45*(2), 168-176. https://doi.org/10.37934/araset.45.2.168176

Timarán Buchely, A., y Timarán Pereira, R. (2023). Minería de datos educativa para descubrir patrones asociados al desempeño académico en competencias genéricas. *Revista Colombiana de Tecnologias de Avanzada (RCTA)*, *2*(38), 87-95. https://doi.org/10.24054/rcta.v2i38.1282

Tsiakmaki, M., Kostopoulos, G., Kotsiantis, S., y Ragos, O. (2019). Implementing autoML in educational data mining for prediction tasks. *Applied Sciences*, *10*(1), 90. https://doi.org/10.3390/app10010090

Universidad de los Andes, Facultad de Economía, Centro de Estudios sobre Desarrollo Económico CEDE. (2014). *Informe Determinantes de la deserción "Informe mensual sobre el soporte técnico y avance del contrato para garantizar la alimentación, consolidación, validación y uso de la información del SPADIES"*. Bogotá D.C.: Colombia. https://www.mineducacion.gov.co/sistemasdeinformacion/1735/articles-254702_Informe_determinantes_desercion.pdf

Xie, Y. (2021). Student performance prediction via attention-based multi-layer long-short term memory. *Journal of Computer and Communications*, *09*(08), 61-79. https://doi.org/10.4236/jcc.2021.98005

Zhang, H., Li, K., y Fu, X. (2010). On pinning control of some typical discrete-time dynamical networks. *Communications in Nonlinear Science and Numerical Simulation*, *15*(2), 182-188. https://doi.org/10.1016/j.cnsns.2009.01.019