Revista Educación, 2025, 49(1), enero-junio

ISSN: 0379-7082 / e-ISSN 2215-2644

Artículo científico de investigación

DOI: http://doi.org/10.15517/revedu.v49i1.61275

Abandono estudiantil en el curso de Matemática General: identificación de variables relevantes para su predicción

Scholar Dropout at General Mathematics subject: identification of relevant variables for its prediction

José Andrey Zamora-Araya
Universidad Nacional
Heredia, Costa Rica
jzamo@una.ac.cr (Correspondencia)
https://orcid.org/0000-0001-6050-5850

Tania Elena Moreira-Mora
Instituto Tecnológico de Costa Rica
Cartago, Costa Rica
tmoreira.costarica@gmail.com
https://orcid.org/0000-0002-8955-0804

Recepción: 2 de septiembre de 2024
Aceptado: 30 de octubre de 2024

¿Cómo citar este artículo?

Zamora-Araya, J. A. y Moreira-Mora, T. E. (2025). Abandono estudiantil en el curso de Matemática General: identificación de variables relevantes para su predicción.*Revista Educación, 49*(1). http://doi.org/10.15517/revedu.v49i1.61275

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

RESUMEN

El objetivo del estudio fue determinar las variables más importantes para la predicción del abandono estudiantil del curso de Matemática General, de la Universidad Nacional de Costa Rica (UNA), considerando el tipo de estudiante y el momento en que se identifica el abandono. Se construyeron seis modelos predictivos (dos grupos estudiantiles en tres momentos diferentes) y en cada modelo se implementaron tres algoritmos de aprendizaje supervisado: Regresión Logística (RL), Random Forest (RF) y XGBoost (XGB). La muestra total se dividió en archivos de entrenamientos, estudiantado que matriculó el curso durante los años 2017 y 2018 y archivos de prueba correspondientes a la matrícula del año 2019. Una vez calibrados los hiperparámetros (validación 10 folds) se identificaron las principales variables asociadas con abandono estudiantil (AE) en el curso de Matemática General de cada modelo con base en la medida de importancia de Gini. No obstante, el rendimiento de los algoritmos oscila entre valores de F1- Score de 0.6251 y 0.7300. Además, se comparó el poder predictivo de los algoritmos en cada modelo por medio de un ANOVA de medidas repetidas con validación cruzada con 10 Folds, y no se encontraron diferencias significativas entre los tres algoritmos en ninguno de los modelos propuestos. Las principales variables asociadas al abandono estudiantil (AE) son de tipo académico como la nota de la prueba de actitud académica (PAA), la nota de colegio y las notas de las pruebas parciales, individuales; como el sexo y la edad de ingreso, económicas; como la beca y el índice de desarrollo social (IDS) e institucionales como el estrato, la edad y especialización del personal docente. Se recomienda asignar al profesorado especializado en Matemática Educativa para impartir los cursos iniciales y el diseño propuesto para tomar decisiones sobre acciones que aumenten la permanencia.

PALABRAS CLAVE: Enseñanza superior, Abandono escolar, Rendimiento escolar, Datos estadísticos, Metodología estadística, Matemáticas.

ABSTRACT

The aim of this study was to determine the most important variables for predicting student dropout from the General Mathematics course (MAT001) of the Universidad Nacional de Costa Rica (UNA), considering the types of students and the time at which dropout takes place. Six predictive models were constructed (two student groups at three different times) and three supervised learning algorithms were implemented in each model: Logistic Regression (LR), Random Forest (RF) and XGBoost (XGB). The total sample was split into training files containing data on students who enrolled in the course during the years 2017 and 2018, and test files with data corresponding to students who enrolled in the year 2019. Once the hyperparameters were fitted (10-fold validation), the main variables associated with student dropout (SD) in the General Mathematics course of each model were identified based on the Gini importance measure; performance of the algorithms ranged from

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

F1-Scores of 0.6251 to 0.7300. In addition, the predictive power of the algorithms in each model were compared by means of a repeated-measures ANOVA with 10-fold cross-validation, and no significant differences were found between the three algorithms in any of the proposed models. The main variables associated with student dropout (SD) were academic, such as grades on the academic attitude test (AAT), high school education grades, and grades on MAT001 tests, student attributes as sex and age at enrollment, economic factors such as scholarships and the Social Development Index (SDI), and institutional factors such as high school educational opportunities that students were exposed to, and the ages and specializations of the teaching staff. Based on the results of this analysis, it is recommended that teachers specialized in Educational Mathematics be assigned to teach initial courses, and to propose designs for decision making about actions that increase permanence.

KEYWORDS: Higher Education, Student Dropout, Student Performance, Statistical Data, Statistical Methodology, Mathematics.

INTRODUCCIÓN

El Abandono Estudiantil (AE) ha sido un fenómeno estudiado desde 1940, inicialmente enfocado en la deserción. A partir de los 90's con los procesos de masificación de la educación terciaria, el problema fue más evidente, ya que el estudiantado que ingresaba a las aulas universitarias era más heterogéneo y estaba presionado por obtener un título universitario o una certificación que brindara mayores posibilidades de incorporación a un mercado laboral competitivo y demandante (Wang et al., 2023).

El AE universitario puede ser analizado a nivel del sistema educativo, de la institución o del plan de estudios en que el estudiantado ingresó, siendo estos dos últimos en los que se concentran la mayor parte de los estudios sobre AE. Más recientemente, se ha despertado el interés por investigar el AE en asignaturas o cursos particulares, entre las que sobresalen aquellos con alto contenido matemático y que suelen darse durante el primer año de universidad. Por ejemplo, en el caso de la Universidad Nacional (UNA) el curso de Matemática General (MAT001) es el que presenta la mayor tasa de abandono pues, si se excluyen los años de pandemia, en promedio menos de la tercera parte del estudiantado logra concluir el curso (Castillo-Sánchez et al., 2020; Zamora-Araya et al., 2020).

Entre las repercusiones del AE a nivel de cursos están aumentar la duración de las carreras e incrementar los costos derivados de la necesidad de ofertar cursos para personas estudiantes repitentes. Además, de poner en dificultades a la administración para satisfacer la demanda de cupos en un escenario de escasez de recursos presupuestarios.

Según el Programa Estado de la Nación (PEN) las tasas de abandono en Costa Rica para la educación superior pública varían entre un 43,4% y 65,3% (Román, 2017). Por eso las instituciones de

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

educación superior se preocupan por plantear estrategias que aumenten los niveles de permanencia y se esfuerzan por ampliar la cobertura y la calidad de los programas académicos ofertados.

En cuanto al uso de técnicas de análisis de datos las investigaciones sobre el tema, en Costa Rica, se han centrado en el análisis de factores asociados al AE en estudios de tipo descriptivos, correlacionales, generalmente asociadas a cohortes de ingreso o carreras específicas (Pascua-Cantero, 2016; Rodríguez-Pineda y Zamora-Araya, 2014), en los cuales se encontró que la edad de ingreso, el sexo, la beca y las variables académicas fueron relevantes para explicar el AE.

Recientemente en Costa Rica se han aplicado técnicas estadísticas más sofisticadas como el análisis de sobrevivencia (Mora, 2016) y la aplicación de modelos predictivos (Solís et al., 2018), pero son pocos los estudios que analizan el fenómeno desde la perspectiva de la predicción del AE a nivel de asignaturas o cursos del área de matemáticas (Zamora-Araya, 2023a), ya que la mayoría se realizan a nivel de carrera o cohorte estudiantil.

En el ámbito internacional hay estudios en cursos de Matemática, pero con un enfoque más explicativo que predictivo. Al respecto, Muñoz-Camacho et al. (2018) realizaron un análisis de regresión logística para determinar las probabilidades de AE en personas estudiantes matriculadas en cursos de Matemática Básica en el periodo 2011 a 2016. Los resultados mostraron tasas de clasificación correcta entre el 59% y 70% según la facultad. Además, encontraron que el matricular la asignatura en un ciclo regular (en lugar de uno intensivo), de libre escolaridad, con grupos mayores a 50 personas, ser hombre y pertenecer a la facultad de Ciencias Económicas aumenta las probabilidades de AE.

Calva et al., (2021) en otro estudio sobre el éxito académico en cursos presenciales de Matemática, medidos en cuanto a aprobación, realizaron un análisis para predecir la aprobación en cursos de nivelación (Matemática, Física, Química y Lenguaje). Los resultados mostraron que el gradiente boosting tuvo el mejor rendimiento y las variables con mayor influencia fueron la nota de la primera prueba, el promedio ponderado, la calificación de postulación, la carrera, la asignatura y el número de materias matriculadas y se obtuvo una precisión del 89.1% y un AUC del 95.5%. Otras variables relevantes sugeridas por el modelo de Regresión Logística fueron la edad, lugar de residencia, estado civil, número de miembros en la familia, tipo de colegio, momento de la matrícula, segmento poblacional (acción afirmativa, mérito territorial, población general), semestre del año (A o B) y jornada (matutina o vespertina).

En una investigación sobre la aprobación de la asignatura de análisis en el estudiantado de las disciplinas MINT (matemática pura, informática, ciencia y tecnología) se compararon los resultados de estas carreras con personas estudiantes que optaron por la carrera de profesor de Matemática y llevaron el curso en el ciclo lectivo 2014-2015 (Kilian et al., 2020). Para ello usaron los algoritmos de regresión logística binaria, regresión logística con regularización de red elástica, máquinas de soporte vectorial y métodos basados en árboles, decisión en las métricas accuracy, precisión, recall, kappa y F1.

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

Los resultados mostraron valores en la métrica F1 de entre 0.60 y 0.73 con una predicción correcta del éxito en el curso para el 75% del estudiantado. Además, los predictores más relevantes fueron: las notas de rendimiento previo específicas de Matemáticas y de rendimiento escolar general, el estudiantado que no está en su primer semestre de matemáticas muestra mayores probabilidades de éxito en comparación con quien lo lleva en el primer semestre y el tipo de modalidad de la educación secundaria. Tanto Kilian et al. (2020) como Calva et al. (2021) no analizaron como variable dependiente el AE, sino variables asociadas con el rendimiento académico como la aprobación y la nota final del curso.

Los estudios internacionales y nacionales reconocen la importancia de utilizar herramientas, como los modelos predictivos, para la detección temprana del estudiantado con altas probabilidades de abandono. El objetivo de esta investigación es determinar las variables relevantes para predecir las probabilidades de AE del estudiado que matricula el curso de Matemática General, de la Universidad Nacional de Costa Rica (UNA), considerando el tipo de estudiante y el momento en que se identifica AE.

La mayor parte de las investigaciones relativas al AE en cursos de Matemática se circunscriben a entornos con algún grado de virtualidad (Kilian et al., 2020; López-Zambrano et al., 2021) debido a la facilidad para recolectar los datos por medio de plataformas informáticas. Por ello, uno de los aportes del estudio es el uso de técnicas de aprendizaje supervisado para analizar el AE en un curso de Matemática a nivel introductorio en la modalidad tradicional (presencial), con muestras relativamente grandes de personas estudiantes, tanto de nuevo ingreso como regulares, que matricularon el curso al menos en una ocasión.

Hacer una distinción entre el estudiantado que matricula por primera vez el curso (nuevo ingreso) y el que lo ha matriculado en más de una ocasión (regulares) es otro aspecto que, a diferencia de otras investigaciones sobre el tema, este estudio considera.

También la evidencia empírica ha mostrado que variables de tipo académico, económico y personal se han asociado con el AE en el curso de Matemática General (Castillo-Sánchez et al., 2020; Zamora-Araya et al., 2020). Sin embargo, existen otras variables relacionadas con los atributos del profesorado, ambientes educativos, características y evaluaciones de los cursos que, junto con variables académicas tradicionales como créditos aprobados, promedio ponderado, evaluaciones diagnósticas también han sido utilizadas para implementar modelos predictivos en el ámbito educativo (Xu et al., 2022). De ahí la importancia de incorporar en este estudio variables institucionales asociadas al AE.

Por último, este documento se organiza con una justificación del problema, los antecedentes nacionales e internacionales del AE, el referente conceptual de las variables asociadas al AE y una descripción del uso de la minería de datos en el ámbito educacional. En la sección metodológica se describe el diseño de la investigación, la población, los criterios de exclusión y el proceso de separación de los archivos para entrenamiento y validación, así como el diseño de los modelos planteados y

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

la estrategia de análisis. Luego se expone una discusión de los resultados y finaliza con la sección de conclusiones y recomendaciones.

Referentes Conceptuales

Definiciones de abandono escolar (AE)

Los estudios sobre abandono escolar (AE) han tratado aspectos conceptuales del término (Cabrera et al., 2006; Tinto, 1989), tipos de abandono (Behr et al., 2020; Munizaga et al., 2018), factores asociados, modelos teóricos para predecir o explicar el fenómeno (Khoushehgir y Sulaimany, 2023; Wang et al., 2023) y de las consecuencias que genera, lo que evidencia el carácter multidimensional y complejo tanto del fenómeno como del término.

Prueba de ello es que no existe claridad sobre el término que se equipara con otros como deserción escolar, absentismo, fracaso escolar o exclusión educativa, aunque no representan el mismo constructo, ni todas las personas usan de la misma forma la terminología para referirse al fenómeno (Zamora-Araya et al., 2023). Esto refleja la falta de acuerdo de la comunidad científica sobre el tema, por lo que las personas investigadoras adoptan una definición propia que responde a sus necesidades e intereses (Valencia et al., 2024). Como lo señala Tinto (1982) “es probable que ninguna definición de abandono capture por completo la complejidad de su aparición en la educación superior” (p. 14).

La definición teórica más próxima al enfoque de esta investigación es la de Bäulke et al. (2022), quienes visualizan el AE como un proceso de toma de decisiones, ya que la decisión final del estudiantado de dejar la universidad puede verse afectada, en parte, por las condiciones de entrada, incluidos los pensamientos e intenciones sobre el AE. Por otra parte, ante la ausencia en la literatura de una definición operativa sobre el AE de un curso o asignatura específica en la modalidad presencial, se recurrió a lo que la *Educational Data Mining* (EDM por sus siglas en inglés) sugiere para la detección temprana del AE, específicamente, al uso de variables provenientes de registros de sistema de aprendizaje electrónico, sistemas de información estudiantiles, encuestas, evaluaciones de cursos y registros de acceso en línea (Xu et al., 2022). De esta manera, se asume en este estudio que una persona abandona el curso de MAT001 si no se presenta a realizar el último examen parcial. Adaptando la clasificación propuesta por Castaño et al. (2004) y Munizaga et al. (2018), se dice que este abandono es precoz si no se presenta a realizar ningún examen parcial, temprano si solo se presenta a realizar la primera prueba parcial y tardío si solo se presenta a las dos primeras pruebas (en el curso se realizaron tres pruebas parciales).

Variables asociadas al AE

El AE puede ser analizado desde diferentes enfoques entre ellos: sociológicos, interaccionistas, organizacionales, psicológicos, económicos e integradores (Lázaro et al., 2020), que han dado como resultado una categorización de las variables que pueden ser resumidas en cuatro determinantes:

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

individuales, socioeconómicos, académicos e institucionales (Guzmán et al., 2021; Zamora-Araya, 2023a). Por tanto, esta investigación asume una categorización de variables más que de factores.

Según la multidimensionalidad y causas diversas del AE, el proyecto Alfa de Gestión Universitaria Integral del Abandono (Alfa-GUIA), actualmente Red GUIA, propuso el siguiente modelo con sus respectivas variables para estudiar la deserción (Proyecto ALFA-GUÍA, 2013).

• Individuales: hace referencia a características personales del estudiantado, como la pérdida de motivación que influye en la deserción y a su vez se asocian con la teoría sobre conductas de logro. Algunas variables asociadas son vocación, dependencia económica, motivación, hábitos de estudio, adaptación a la vida universitaria, entre otras.

• Académicos: relacionados con la institución de procedencia, puntaje en las pruebas previas al ingreso a la universidad, conocimientos previos, hábitos de estudio, número de créditos matriculados y demás variables que inciden en el rendimiento y a su vez en el abandono escolar.

• Económicos: incorpora variables relacionadas con aspectos monetarios y uso de recursos económicos como ingreso familiar y personal, capacidad para financiar los estudios y satisfacción al concluir la carrera.

• Culturales: vinculados con las creencias, valores y prácticas que forman parte del contexto cultural de la persona estudiante que puede incidir en su estabilidad emocional y motivación.

• Institucionales: asociados con variables relacionadas con la beca, la calidad de la docencia, atención psicosocial, junto con las características e integración del personal docente con el estudiantado, entre otros.

AE y Minería de Datos en Educación

Durante la última década los estudios sobre AE han incorporado modelos predictivos como herramienta para el análisis del éxito académico estudiantil. Esta tendencia de usar técnicas de minería de datos y Machine Learning (ML) a problemas relacionados con actividades de aprendizaje en ambientes educacionales se enmarca en la denominada EDM.

Entre las variables más importantes para predecir el rendimiento académico en el contexto de EDM están la nota media acumulada junto con las evaluaciones internas y, en menor medida, las variables sociodemográficas del estudiantado (sexo, estatus socioeconómico, nivel de pobreza, edad, beca, ocupación de los padres), las actividades extracurriculares (deportivas, artísticas, culturales), los antecedentes de la escuela secundaria (notas, pruebas de admisión) y la red de interacción social como uso de celular, actividad en plataformas y redes sociales (Albreiki et al., 2021; Khan y Ghosh, 2021; Khoushehgir y Sulaimany, 2023).

Las técnicas de ML se han aplicado en cursos en línea y presenciales, donde la variable dependiente suele ser la nota final de la asignatura o el nivel de aprobación (reprobación) en el curso, en lugar del AE (Alvarado y Zambrano, 2020; Calva et al., 2021).

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

En lo que respecta a cursos del área de Ciencias y Matemática, los estudios utilizan la información de las asignaturas, ya sea como variables independientes que ayudan a predecir la titulación (Opazo et al., 2021) o como variables dependientes en donde se busca predecir las notas, el nivel de aprobación y, en menor medida, la permanencia o AE en el curso (Alvarado y Zambrano, 2020; Kilian et al., 2020). Al respecto Shin y Shim (2021) señalan que los estudios prestan más atención a la identificación de factores importantes que influyen en el rendimiento del estudiantado en Matemáticas o Ciencias que en la predicción de su rendimiento, pues estos son relativamente pocos y los existentes recolectan datos por medio de plataformas en cursos en línea para obtener información sobre lo que hace o no hace el estudiantado.

METODOLOGÍA

Tipo de Investigación

Se plantea un estudio de cohortes de tipo correlacional predictivo por su enfoque en el estudio de eventos que ocurrirán en el futuro (Hernández et al., 2014). Para realizar la predicción se utilizaron tres algoritmos de aprendizaje supervisado: Regresión Logística (RL), Random Forest (RF) y XGBoost (XGB).

Participantes

Se consideraron tres cohortes de personas estudiantes que matricularon el curso de MAT001 en la UNA entre los años 2017 y 2019, es decir, toda la población estudiantil que, durante los años 2017, 2018 y 2019 matriculó al menos en una ocasión el curso MAT001, ya sea en el primer o segundo ciclo lectivo de cada año. El archivo de datos contenía 5906 registros, pero al aplicar los criterios de exclusión se desestimaron 91 casos para un total de 5815 registros disponibles para análisis.

Criterios de exclusión

La exclusión de personas estudiantes se basó en los siguientes criterios: (a) realizaron retiro justificado, (b) no pudieron ser ubicados o no tenían identificación, (c) sin registro de notas parciales o registros incompletos, (d) decidieron realizar el curso por suficiencia y (e) sin información completa del proceso de admisión, por ejemplo, nota de admisión, estrato y nota de colegio, entre otras.

Con este último criterio quedaron excluidas las personas estudiantes que ingresaron a la UNA por convenios o modalidades y las que ingresaron antes de 2009, ya que, desde este año, hubo un cambio en el sistema de admisión que incluyó el proceso de estratificación.

Recolección de la información

Los datos fueron suministrados por dos instancias: la coordinación de Cursos de servicio de la Escuela de Matemática con registros del cuerpo docente y de las notas para la asignatura para 2017, 2018 y 2019 y el Departamento de Registro, que suministró la información relacionada con el proceso

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

de admisión, rendimiento académico y variables socioeconómicas, lo cual hizo mediante un oficio, donde se estipula la confidencialidad y el manejo de la información acorde con los criterios de la instrucción UNA-AJ-DICT-17-2020, emitida por el Departamento de Asesoría Jurídica de la UNA, en concordancia con la ley 8968 de protección de sus datos personales (Asamblea Legislativa, 2011).

Proceso de separación en archivos de entrenamiento y validación

Lo usual en los modelos de Machine Learning es separar el conjunto de datos en dos: uno de entrenamiento y otro de prueba. Lo frecuente es tomar aleatoriamente un porcentaje del total de datos disponibles para entrenamiento (de 70% a 90%) y el resto como datos de prueba.

En esta investigación el archivo de entrenamiento lo constituyó el estudiantado que matriculó el curso durante 2017 y 2018 y la matrícula del 2019 se dejó como archivo de prueba. Esto hace que ambos conjuntos (entrenamiento y validación) no fueron seleccionados aleatoriamente, ya que se toma el año como criterio de separación.

La forma de dividir los conjuntos de entrenamiento y prueba representa la situación realista, donde se aplicaría la predicción. Esto significa predecir el AE en el curso MAT001 de estudiantes a partir de la información disponible de los ciclos anteriores, los registros universitarios y conforme avanza el semestre. La forma de dividir los archivos es una fortaleza al emular los escenarios reales donde la propuesta podría llevarse a la práctica, pues toma como punto de partida la información existente de estudiantes en las bases institucionales para predecir el fenómeno de interés, en este caso el AE, en el estudiantado como otros estudios (Kilian et al., 2020).

Descripción del diseño

Para el diseño se dividió a la población estudiantil en dos grupos según condición de ingreso. El primero está constituido por el estudiantado de nuevo ingreso de las cohortes 2017, 2018 y 2019 que matricularon el curso MAT001 durante su primer semestre de universidad. Para esta población solo se cuenta con información proveniente de variables asociadas al proceso de admisión y al curso; como las notas de las pruebas parciales, así como las institucionales relativas a las características del profesorado de la asignatura. El segundo grupo lo integran estudiantes regulares, que matricularon el curso MAT001 entre el primer ciclo de 2017 y el segundo de 2019, pero que tenían al menos un semestre de clases en la universidad, incluyendo al estudiantado que ingresó en años anteriores a 2017. Además de la información referente al proceso de admisión y del curso, este grupo cuenta con un historial académico universitario que amplía la cantidad de variables disponibles para la predicción, como el promedio ponderado o la cantidad de materias abandonadas del semestre anterior.

Para el diseño también se consideró la clasificación temporal propuesta por Castaño et al. (2004):

1. Precoz: representa al estudiantado que, habiendo realizado todos los trámites para admisión, no concreta su matrícula.

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

2. Temprano: considera al estudiantado que se retira de los estudios durante los primeros cuatro semestres de la carrera (dependiendo del estudio el número de semestres puede variar).

3. Tardío: contempla al estudiantado que abandona los estudios en un período posterior a lo establecido por la deserción temprana.

Considerando esta clasificación temporal se definieron las siguientes categorías utilizando el número de pruebas realizadas como un marcador de la dimensión temporal del AE: (1) precoz para personas estudiantes que matricularon el curso, pero no realizaron pruebas parciales, pues la mayoría de los casos ni siquiera asistieron a clases; (2) temprano para quienes realizaron la primera prueba parcial, pero no realizaron la segunda ni la tercera y (3) tardío para quienes realizaron las dos primeras pruebas parciales, pero no la tercera.

Al considerar de manera conjunta ambas poblaciones (nuevo ingreso y regulares) y las categorías temporales (abandono precoz, abandono temprano y tardío) se diseñaron seis modelos para analizar el AE: (1) nuevo ingreso - ninguna prueba parcial, (2) nuevo ingreso - solo primera prueba parcial, (3) nuevo ingreso - dos primeras pruebas parciales, (4) regular- ninguna prueba parcial, (5) regular - solo primera prueba parcial, (6) regular - dos primeras pruebas parciales. Cada uno difiere por sus particularidades y variables disponibles para la predicción.

La aplicación de este diseño constituye un aporte metodológico a los estudios sobre AE, pues por lo general las investigaciones no realizan divisiones de este tipo para realizar la predicción, o bien, consideran otras características asociadas al área de conocimiento de la carrera, el sexo o afiliación institucional (Albreiki et al., 2021; Khan y Ghosh; 2021; Kilian et al., 2020).

Además, en los seis modelos se ejecutaron tres algoritmos diferentes de aprendizaje supervisado: (1) Regresión Logística (RL), (2) Random Forest (RF) y (3) XGBoost (XGB). La razón de usar estos algoritmos es que se desea comparar los resultados de la RL, considerado un algoritmo de caja blanca por su ventaja de conocer en detalle el proceso de obtención de resultados y la facilidad para interpretarlos, con los otros dos algoritmos de caja negra, uno tipo *bagging* y otro tipo *boosting* que, a pesar de su mayor complejidad tanto en términos de interpretación como de procesamiento, tienen el potencial de brindar un mejor rendimiento predictivo. Asimismo, tanto para comparar el rendimiento de los algoritmos como para ajustar sus hiperparámetros, se utilizó el método de validación cruzada 10 *folds*, la cual internamente realiza una selección aleatoria de instancias para hacer la validación en cada *fold*.

En la Tabla 1 se muestra la cantidad de personas estudiantes en cada archivo de entrenamiento y prueba por modelo.

Tabla 1.

Número de registros y porcentaje de AE para cada modelo predictivo del curso MAT001 según número de pruebas y condición de ingreso para los archivos de entrenamiento y prueba 2017-2019

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

Cantidad de pruebas incluidas en los medios	Nuevo Ingreso				Regular			
	Archivo de entre-namiento 2017 y 2018	% AE	Archivo de prue-ba 2019	% AE	Archivo de entre-namiento 2017 y 2018	% AE	Archivo de prue-ba 2019	% AE
Ninguna Prueba Parcial	1348	51.93	771	44.2	1581	51.6	936	45.9
Solo Prime-ra Prueba Parcial	1186	45.36	669	35.7	1337	42.8	794	36.3
Primera y Segunda Prueba	1003	35.39	599	28.2	1056	27.6	652	22.4

Fuente: Elaboración propia.

Para cada modelo se dividió el archivo principal en dos, uno para entrenar el modelo y otro para validarlo. Se puede observar que los porcentajes de AE en los archivos de validación y entrenamiento son similares, a pesar de ser muestras de años diferentes. Se seleccionó la métrica F1 como medida de comparación para el rendimiento de los algoritmos por ser la que brinda un mejor balance entre la precisión y el recall. Para realizar los análisis se utilizó el software estadístico R versión 4.2.1 y los paquetes boruta para la selección de variables y caret y tidymodels para el ajuste de los modelos predictivos y generar la medida de importancia de Gini para cada variable en el caso de RF y XGB.

Descripción de la estrategia de análisis de la información

La muestra estuvo constituida por quienes matricularon el curso durante los ciclos lectivos que van desde el primer ciclo de 2017 al segundo ciclo de 2019. Dado que el objetivo es predecir el AE en el curso, estos casos representan la muestra disponible de todas las posibles personas que se matricularon o podrían matricular el curso en el futuro.

El primer paso consistió en la limpieza del archivo de datos suministrado por el Departamento de Registro y la Escuela de Matemática. Luego se planteó el diseño de seis modelos que se evaluaron por medio de la validación cruzada k *folds* (k = 10).

El segundo paso fue elegir las variables para los modelos, mediante un análisis descriptivo y correlacional donde se excluyeron aquellas variables con información redundante, como por ejemplo la nota de admisión compuesta por la nota de colegio y la nota en la PAA. Para los casos de variables

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

altamente correlacionadas ($r > 0.80$), o que presentaron una relación lineal entre ellas, se seleccionó una. Las variables con insuficiente cantidad de casos por categoría también se excluyeron, por ejemplo, la nacionalidad por afectar las estimaciones de los modelos.

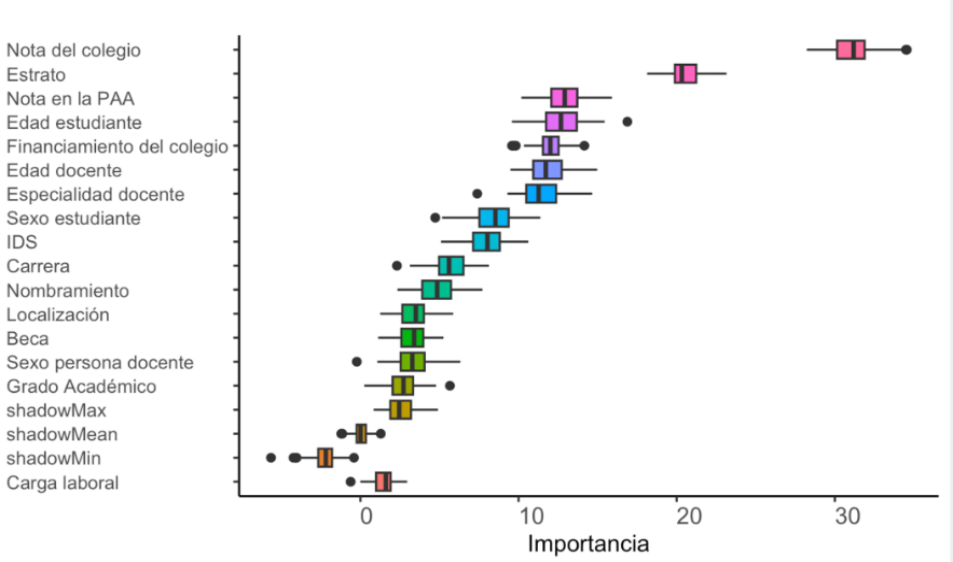
Luego se aplicó el algoritmo de selección de características boruta para seleccionar las variables con mayor potencial de predicción para cada modelo; así se redujo el total de variables de 148 a 18 para estudiantes de nuevo ingreso y a 22 para los regulares.

A modo de ejemplo, la Figura 1 muestra la selección hecha por boruta para el modelo 1, el cual será utilizado para ejemplificar cada fase del procedimiento.

Los datos y el código necesarios para reproducir los demás modelos pueden ser consultados en el sitio https://acortar.link/qQNIyv.

Figura 1.

UNA: Variables seleccionadas por el algoritmo Boruta para el estudiantado de primer ingreso sin evaluaciones parciales



Fuente: Elaboración propia.

El tercer paso, una vez seleccionadas las variables para cada modelo, consistió en usar una validación cruzada de 10 *folds* para ajustar los hiperparámetros de los algoritmos RF y XGB (la RL no tiene hiperparámetros). Por ejemplo, para el RF se ajustaron los hiperparámetros de número de árboles incluidos en el modelo, la profundidad máxima del árbol y el número mínimo de observaciones por nodo, para el XGB los hiperparámetros eta, gamma, max_dept y min_child_weihgt. Tal procedimiento se aplicó en cada algoritmo de los seis modelos.

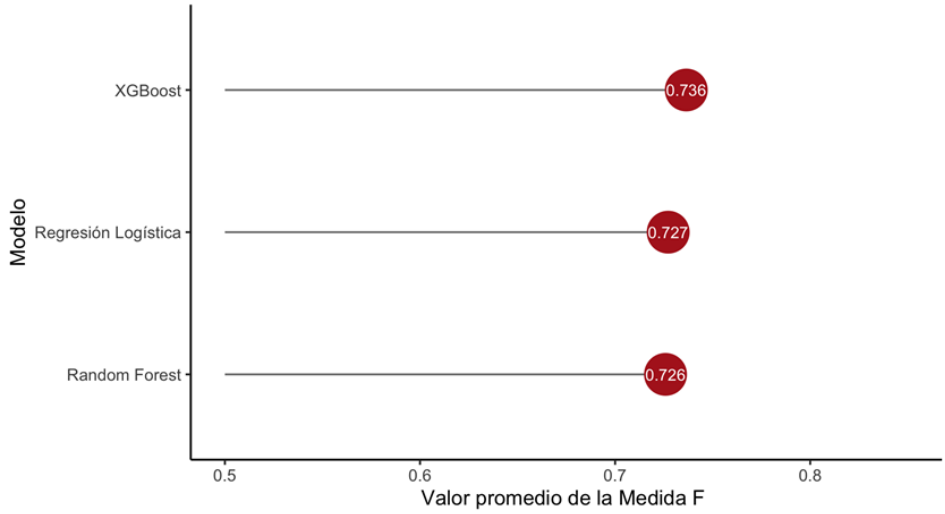
Como cuarto paso, una vez determinados los parámetros óptimos para cada algoritmo en cada modelo, se realiza una nueva validación cruzada 10 *folds* para entrenar los tres algoritmos en cada

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

uno de los seis modelos, teniendo el cuidado de colocar una semilla aleatoria que permita reproducir los resultados y garantizar que cada algoritmo usa el mismo *fold* en cada una de las 10 iteraciones. Esto se hace con el conjunto de entrenamiento (cohortes 2017 y 2018). A modo de ejemplo, la Figura 2 muestra el promedio de la medida de comparación (F1 score) para cada algoritmo en el modelo 1 (estudiantes de nuevo ingreso, ninguna prueba parcial).

Figura 2.

UNA: Comparación de los algoritmos de Regresión Logística, Random Forest y Xgboost por validación cruzada 10 *folds* para el estudiantado de primer ingreso sin evaluaciones parciales



Fuente: Elaboración propia.

En el quinto paso, por medio de un ANOVA de medidas repetidas (cada iteración es una réplica), se evalúan los resultados para determinar si existe evidencia a favor de algún algoritmo en cuanto a rendimiento predictivo para cada modelo. Antes de aplicar el ANOVA se verificó el supuesto de normalidad que se comprobó con prueba de Anderson Darling que brindó valores p mayores a 0.05 en todos los casos. Al aplicar el ANOVA, no se rechazó la hipótesis nula de igualdad de promedios (valor $p > 0.05$), por lo que no hay evidencia a favor de alguno de los algoritmos en la métrica seleccionada. A modo de ejemplo, la Figura 3 muestra el diagrama de cajas para el ANOVA del modelo 1.

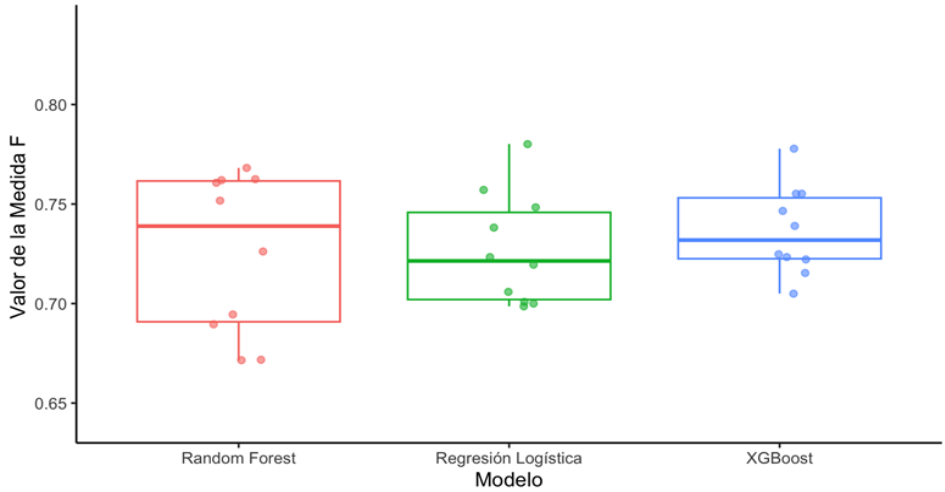
En el paso 6, ya que no hay evidencia para decir que algún modelo tenga un rendimiento significativamente mejor según la métrica F1, se evalúan los tres algoritmos en los datos de prueba correspondientes (datos nunca vistos por el modelo) en los seis modelos.

Por último, para determinar las variables más relevantes en la predicción del AE en cada algoritmo de caja negra, se utilizó la medida de importancia de Gini y la medida de *odds ratio* en el caso de la regresión logística. La Figura 4, Figura 5 y Figura 6 muestran la importancia relativa de cada variable para cada algoritmo del modelo 1.

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

Figura 3.

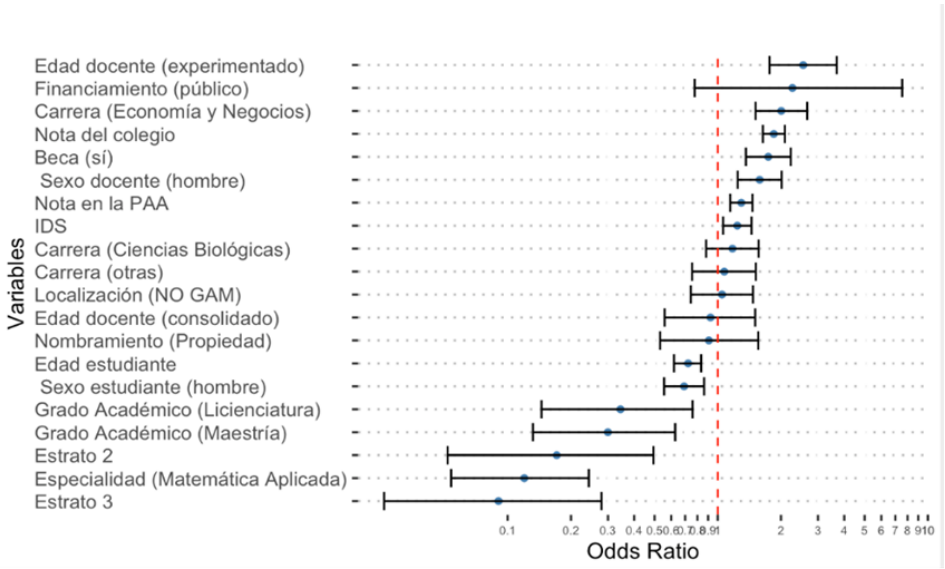
UNA: Diagrama de cajas que representa el ANOVA que compara los algoritmos de Regresión Logística, Random Forest y XGBoost por validación cruzada 10 *folds* para el estudiantado de primer ingreso sin evaluaciones parciales



Fuente: Elaboración propia.

Figura 4.

UNA: Importancia de las variables para el estudiantado de primer ingreso según el algoritmo de Regresión Logística sin evaluaciones parciales

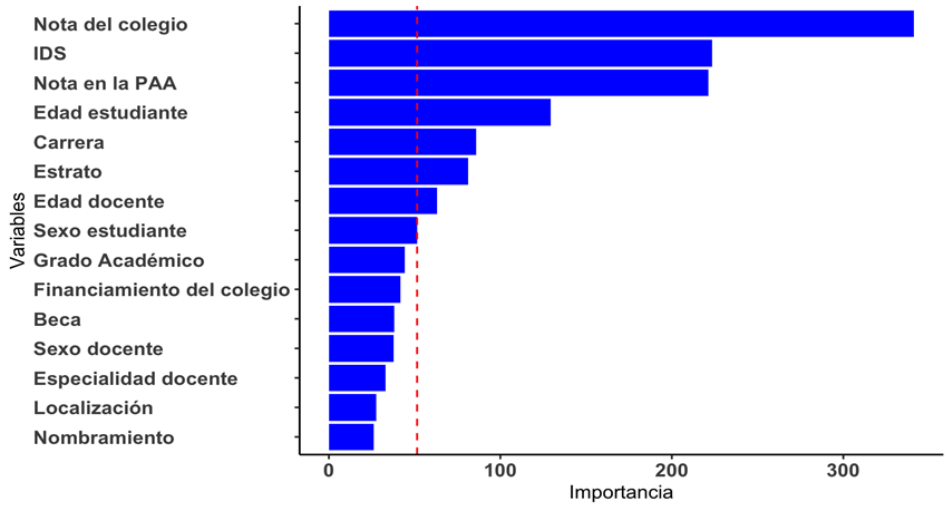


Fuente: Zamora-Araya (2023b) p. 143.

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

Figura 5.

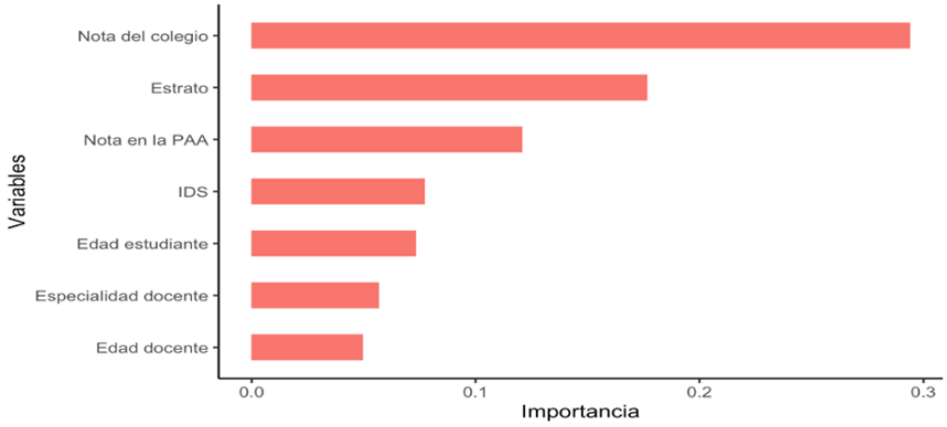
UNA: Importancia de las variables para el estudiantado de primer ingreso según el algoritmo de Random Forest sin evaluaciones parciales



Fuente: Zamora-Araya (2023b) p. 144.

Figura 6.

UNA: Importancia de las variables para el estudiantado de primer ingreso según el algoritmo XG-Boost sin evaluaciones parciales



Fuente: Zamora-Araya (2023b) p. 144.

RESULTADOS Y DISCUSIÓN

Uno de los principales resultados fueron los valores de la métrica de comparación F1 obtenidos al evaluar los algoritmos en los archivos de prueba propuestos en cada uno de los modelos, que se resumen en la Tabla 2.

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

Como se aprecia, los modelos, tienen un poder predictivo similar y no se puede afirmar la superioridad de ninguno de ellos. Sin embargo, el Random Forest y la Regresión Logística mostraron valores ligeramente mayores en la mayoría de los modelos.

La importancia de las características en cada modelo varió, como se puede observar en la Tabla 3. Por ejemplo, en el primer modelo las variables identificadas como relevantes en los dos algoritmos con mejor rendimiento (RL y RF) fueron: las notas de colegio y PAA, el estrato y la edad tanto del estudiantado como del profesorado; otras variables menos relevantes fueron el IDS, la carrera, la beca, el sexo del estudiantado y la especialidad docente.

En varios estudios se ha confirmado la relación de la nota de colegio y la PAA con el rendimiento y permanencia de los estudiantes en el curso MAT001 (Castillo-Sánchez et al., 2020; Zamora-Araya et al., 2020). También la beca ha mostrado ser un factor protector del AE, al igual que el IDS (Zamora-Araya et al., 2020).

En cuanto a la edad de ingreso y el sexo del estudiantado los modelos identificaron que a mayor edad y ser hombre aumenta las probabilidades de AE, variables que suelen asociarse con rezagos en niveles educativos previos o presión por incorporase al mundo laboral a temprana edad (Lázaro et al., 2020; Mora, 2016).

Tabla 2.

Valor de la métrica F1 para cada modelo según tipo de estudiante y cantidad de pruebas parciales en el periodo 2019

		Algoritmo		
	Modelo según variables incluidas	Regresión logística	Random Forest	XGBoost
Nuevo Ingreso	Solo variables de admisión	0.62	0.61	0.63
	Incluye primera prueba parcial	0.70	0.70	0.65
	Incluye ambas pruebas parciales	0.71	0.71	0.70
Regulares	Solo variables de admisión	0.66	0.68	0.67
	Incluye Primera prueba parcial	0.73	0.72	0.65
	Incluye ambas pruebas parciales	0.70	0.70	0.68

Fuente: Elaboración propia.

Tabla 3.

UNA: Variables consideradas relevantes en cada algoritmo según el modelo propuesto

Modelo	Regresión logística	Random Forest	XGBoost

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

1. Ninguna prueba parcial (nuevo ingreso)	Edad docente Carrera Nota de colegio Nota en la PAA Beca Sexo (Docente) Edad (Estudiante) Sexo (Estudiante) Grado académico (profesorado) Estrato 2 Estrato 3 Especialidad (profesorado)	Nota de colegio IDS Nota en la PAA Edad (Estudiante) Carrera Estrato Edad (Docente) Sexo (Estudiante)	Nota de colegio Estrato Nota en la PAA IDS Edad (Estudiante) Especialidad (docente) Edad (Docente)
2. Primera prueba parcial (nuevo ingreso)	Nota parcial 1 Edad docente Carrera Beca Nota de colegio Sexo (Estudiante) Edad (Estudiante) Grado Académico (profesorado) Especialidad (docente)	Nota parcial 1 Nota de colegio Edad (Estudiante) Estrato IDS Especialidad (docente)	Nota parcial 1 Nota de colegio Edad (Estudiante)
3. Ambas pruebas parciales (nuevo ingreso)	Nota parcial 1 Nota parcial 2 Edad docente Especialidad (docente)	Nota parcial 1 Nota parcial 2 Nota en la PAA IDS Edad (Estudiante) Estrato	Nota parcial 1 Nota parcial 2 IDS Edad (Estudiante) Carrera
4. Ninguna prueba parcial (Regular)	Matricular Beca Nota de colegio Nota en la PAA Promedio ponderado Edad (Estudiante) Sexo (Estudiante) Nombramiento (docente) Materias desertadas	Promedio ponderado Materias desertadas Nota de colegio Nota en la PAA Edad (Estudiante) Matrícula Materias aprobadas Carrera	Materias desertadas Promedio ponderado Nota de colegio Edad (Estudiante) Matrícula Nombramiento Materias aprobadas Carrera Beca Sexo (Estudiante)

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

5. Primera prueba parcial (regular)	Nota parcial 1 Edad (Estudiante) Materias desertadas Sexo (Estudiante) Matricula	Nota parcial 1 Materias desertadas Promedio ponderado Nota de colegio Matrícula Edad (Estudiante) Nota en la PAA	Nota parcial 1 Promedio ponderado Matrícula Materias desertadas Edad (Estudiante) Sexo (Estudiante) Nota de colegio
6. Ambas pruebas parciales (regular)	Nota parcial 1 Nota parcial 2 Materias desertadas Nota en la PAA	Nota parcial 1 Nota parcial 2 Promedio ponderado Nota en la PAA Nota de colegio Materias desertadas Edad (Estudiante)	Nota parcial 1 Nota parcial 2 Promedio ponderado Matrícula Nota en la PAA Materias desertadas Edad (Estudiante) Beca Nota de colegio

Nota. El nombre de las variables con mayores puntajes de importancia en cada algoritmo se destacaron en negrita.

Fuente: Elaboración propia.

El estrato es una variable relevante en el contexto de la UNA, pues representa una aproximación a las condiciones materiales y oportunidades educativas del estudiantado en su etapa de colegio. El estrato 1 está constituido por colegios privados, científicos y bilingües experimentales, el estrato 2 por colegios académicos y técnicos públicos y el estrato 3 por las demás modalidades educativas como colegios nocturnos, colegios rurales, educación a distancia, etc. (Aguilar-Fernández et al., 2024). En los resultados se identificó que el estudiantado del estrato 3, aquellos con menores oportunidades educativas, son los más propensos al AE, lo que confirma la vulnerabilidad de este grupo estudiantil (Rodríguez-Pineda y Zamora-Araya, 2014; Zamora-Araya et al., 2023).

En el caso de las personas estudiantes regulares también resultaron relevantes las variables académicas como la nota de colegio, la nota en la PAA y otras provenientes del historial académico como la cantidad de materias reprobadas, promedio ponderado (ambas del ciclo anterior) y la cantidad de veces que matriculó el curso (a mayor cantidad, mayor probabilidad de AE) y en menor medida la edad del estudiantado y la beca. Las variables más importantes fueron sus notas en los modelos con ambas pruebas parciales.

De las variables institucionales, el recibir clases con docentes experimentados (más de 45 años) y con un postgrado en el área de educación matemática disminuye la probabilidad de AE en comparación con el profesorado más joven o especialistas en áreas de matemática aplicada y pura (ver Figura

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

4). Esto comprueba la relevancia de la especialización en matemática educativa y la capacitación del profesorado en didáctica y evaluación, puesto que las tasas más altas de AE se presentan en los cursos iniciales (Munizaga et al., 2018; Opazo et al., 2021).

En resumen, en la mayoría de los modelos la nota del colegio, la nota en la PAA y la edad del estudiantado se señalan como variables a considerar, independientemente de si se trata de estudiantes de nuevo ingreso o regulares, o se incluyan o no las notas de las pruebas parciales; lo que denota la importancia de estas variables asociadas con el nivel de conocimientos previos, no sólo en Matemática sino en otras áreas del conocimiento.

CONCLUSIONES

El propósito de la investigación se enfocó en identificar las variables relevantes para predecir las probabilidades de AE del estudiado que matriculó el curso de Matemática General, considerando el tipo de estudiante y el momento en que se identifica AE.

Las variables más importantes en los modelos predictivos fueron las académicas, junto con algunas características del profesorado, el estrato (y el colegio de procedencia), la edad de ingreso y el sexo de la persona estudiante que también resultaron predictores del AE.

Por ejemplo, en los modelos para las personas estudiantes de nuevo ingreso tanto el estrato (condiciones materiales y oportunidades educativas) como la especialidad docente resultaron significativas en la mayoría de los algoritmos, aun incorporando las notas de las pruebas parciales. Este hallazgo confirma que la población más vulnerable tiene más probabilidades de AE. De ahí, la importancia de los programas institucionales y la asignación de profesionales del área de matemática educativa, que diseñen actividades novedosas para mejorar los resultados académicos.

Otro ejemplo sería utilizar el modelo para el estudiantado de nuevo ingreso que no incorpora pruebas parciales para plantear alguna estrategia preventiva que disminuya el AE antes de iniciar el ciclo lectivo. El modelo sugiere que las variables más relevantes son las notas de colegio, la nota en la PAA, el estrato, el IDS y la edad del estudiantado; tomando como base esta información, se pueden diseñar intervenciones que tomen en cuenta tales características.

Si se pretende apoyar a las personas estudiantes repitentes antes de comenzar el ciclo lectivo, los análisis muestran que variables académicas como el promedio ponderado, el número de materias aprobadas y desertadas en el ciclo previo, la edad del estudiantado y las características del profesorado son variables relevantes. Tales características podrían ser utilizadas para conformar grupos estudiantiles que puedan ser atendidos por el profesorado idóneo, según los modelos, que permita mejorar los niveles de permanencia y rendimiento en el curso MAT001.

Dada la relevancia de las evaluaciones parciales, en particular la nota de la primera prueba parcial se recomienda a la cátedra del curso proponer acciones evaluativas y metodológicas que favorezcan un

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

mayor rendimiento en estas pruebas. Algunas actividades que podrían ejecutarse son el aumento en el número de pruebas parciales y la posibilidad al estudiantado de reponer la primera prueba parcial del curso, dada la repercusión que tiene sobre el AE.

Asimismo, futuros estudios de corte cualitativo o mixto podrían profundizar sobre las variables relevantes para el AE en cursos de Matemática, como las características del profesorado o el grado de dominio de conocimientos previos. También pueden implementar otros algoritmos, probar los aquí descritos en otros cursos y explorar la relación con otros constructos asociados con el AE como el compromiso estudiantil, la motivación, el sentido de pertenencia institucional o el proceso de ajuste a la vida universitaria. Para los resultados del presente estudio, es recomendable utilizar el algoritmo de regresión logística, ya que su interpretación es más simple y para los usuarios de los modelos resultará más clara la relación entre el AE y las variables independientes en cada modelo.

REFERENCIAS

Aguilar-Fernández, E., Zamora-Araya, J. A. y Rodríguez-Pineda, M. (2024). Análisis de correspondencia simple para estudiar la relación entre factores del abandono escolar y el estrato del colegio de procedencia en la Universidad Nacional de Costa Rica. *Revista Educación, 48*(2), 1-20. https://doi.org/10.15517/revedu.v48i2.58519

Albreiki, B., Zaki, N. y Alashwal, H. (2021). Una revisión sistemática de la literatura sobre la predicción del rendimiento de los estudiantes utilizando técnicas de aprendizaje automático. *Ciencias de la Educación, 11*(9), 1-27. https://doi.org/10.3390/educsci11090552

Alvarado, O. y Zambrano, S. M. (2020). *Modelo predictivo para determinar el fracaso de matemáticas en grado 11 usando machine learning* [Proyecto de grado, Universidad Distrital Francisco José Caldas]. Repositorio Institucional Universidad Distrital - RIUD. https://repository.udistrital.edu.co/bitstream/handle/11349/25365/OmarAlvaradoSantosZambrano2020.pdf?sequence=1&isAllowed=y

Asamblea Legislativa. (2011). *Ley 8968 Protección de la persona frente al tratamiento de sus datos personales*. http://www.pgrweb.go.cr/scij/Busqueda/Normativa/Normas/nrm_texto_completo.aspx?param1=NRTC&nValor1=1&nValor2=70975&nValor3=85989

Bäulke, L., Grunschel, C. y Dresel, M. (2022). Deserción estudiantil en la universidad: Una visión por fases sobre el abandono de los estudios y el cambio de carrera. *Revista Europea de Psicología de la Educación, 37*(3), 853-876. https://doi.org/10.1007/s10212-021-00557-x

Behr, A., Giese, M., Teguim-Kamdjou, H. D. y Theune, K. (2020). Abandono de la universidad: una revisión de la literatura. *Revista de Educación, 8*(2), 614-652. https://doi.org/10.1002/rev3.3202

Cabrera, J. T., Álvarez, P. y González, M. (2006). El problema del abandono de los estudios universitarios. *RELIEVE. Revista Electrónica de Investigación y Evaluación Educativa, 12*(2), 171-203. https://doi.org/10.7203/relieve.12.2.4226

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

Calva, K., Flores, M., Porras, H. y Cabezas-Martínez, A. (2021). Modelo de predicción del rendimiento académico para el curso de nivelación de la escuela politécnica nacional a partir de un modelo de aprendizaje supervisado. *Latin-American Journal of Computing, 8*(2), 58-71. https://doi.org/10.5281/zenodo.5770905

Castaño, E., Gallón, S., Gómez, K. y Vásquez, J. (2004). Deserción estudiantil universitaria: Una aplicación de modelos de duración. *Lecturas de Economía,* (60), 39-65. https://doi.org/10.17533/udea.le.n60a2707

Castillo-Sánchez, M., Gamboa-Araya, R. y Hidalgo-Mora, R. (2020). Factores que influyen en la deserción y reprobación de estudiantes de un curso universitario de matemáticas. *Uniciencia, 34*(1), 219-245. http://dx.doi.org/10.15359/ru.34-1.13

Guzmán, A., Barragán, S. y Cala Vitery, F. (2021). Deserción escolar en la educación superior rural: una revisión sistemática. *Fronteras de la educación, 6*, 1-14. https://doi.org/10.3389/feduc.2021.727833

Hernández, R., Fernández, C. y Batista, M. P. (2014). *Metodología de la investigación* (6ta ed.). Mc Graw Hill.

Khan, A. y Ghosh, S. K. (2021). Análisis y predicción del rendimiento de los estudiantes en el aprendizaje en el aula: una revisión de los estudios de minería de datos educativos. *Educación y Tecnologías de la Información, 26*, 205-240. https://doi.org/10.1007/s10639-020-10230-3

Khoushehgir, F. y Sulaimany, S. (2023). Negative link prediction to reduce dropout in massive open online courses [Predicción de enlaces negativos para reducir la deserción en cursos masivos abiertos en línea]. *Education and Information Technologies,* 1-20. https://doi.org/10.1007/s10639-023-11597-9

Kilian, P., Loose, F. y Kelava, A. (2020). Predecir el éxito de los estudiantes de matemáticas en la fase inicial de la universidad con información dispersa utilizando enfoques de aprendizaje estadístico. *Fronteras de la Educación, 5*, 1-16. https://doi.org/10.3389/feduc.2020.502698

Lázaro, N., Callejas, Z. y Griol, D. (2020). Factores que inciden en la deserción estudiantil en carreras de perfil ingeniería informática. *Revista Fuentes, 22*(1), 105-126. https://hdl.handle.net/11162/200868

López-Zambrano, J., Lara-Torralbo, J. A. y Romero-Morales, C. (2021). Predicción temprana del rendimiento del aprendizaje de los estudiantes a través de la minería de datos: una revisión sistemática. *Psicotema, 33*(3), 456-465. https://reunido.uniovi.es/index.php/PST/article/view/17117

Mora, Y. (2016). *Estudio longitudinal de la deserción universitaria en el Instituto Tecnológico de Costa Rica* [Tesis de maestría, Universidad de Costa Rica]. Repositorio SIBDI. https://repositorio.sibdi.ucr.ac.cr/handle/123456789/22144

Munizaga, F., Cifuentes, M. B. y Beltrán, A. (2018). Retención y abandono estudiantil en la educación superior universitaria en América Latina y el caribe: Una revisión sistemática. *Education Policy Analysis Archives, 26*(61), 1-36. http://dx.doi.org/10.14507/epaa.26.3348

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

Muñoz-Camacho, S. V., Gallardo, T., Muñoz-Bravo, M. y Muñoz-Bravo, C. A. (2018). Probabilidad de deserción estudiantil en cursos de matemáticas básicas en programas profesionales de la Universidad de los Andes-Venezuela. *Formación Universitaria, 11*(4), 33-42. http://dx.doi.org/10.4067/S0718-50062018000400033

Opazo, D., Moreno, S., Álvarez-Miranda, E. y Pereira, J. (2021). Análisis de la deserción universitaria de primer año a través de modelos de aprendizaje automático: Una comparación entre universidades. *Matemáticas, 9*(20), 1-27. https://doi.org/10.3390/math9202599

Pascua-Cantero, P. M. (2016). Factores relacionados con la deserción en el primer y segundo año de estudio en la carrera de Enseñanza de la Matemática de la Universidad Nacional de Costa Rica. *Revista Electrónica Educare, 20*(1), 96-118. http://dx.doi.org/10.15359/ree.20-1.5

Proyecto ALFA-GUÍA. (2013). *Marco conceptual sobre el abandono. Hacia la gestión colectiva de un marco conceptual para analizar, predecir, evaluar y atender el abandono estudiantil en la educación superior*. https://www.scribd.com/document/261888622/Marco-Conceptual-sobre-el-Abandono-pdf

Rodríguez-Pineda, M. y Zamora-Araya, J. A. (2014). *Análisis de la deserción en la Universidad Nacional desde una perspectiva longitudinal. Quinto informe estado de la educación.* Programa Estado de la Nación. https://doi.org/10.13140/RG.2.2.30416.66569

Román, M. (2017). Capítulo 5: La evolución de la educación superior. En *Sexto informe estado de la educación* (pp. 241-308). Programa Estado de la Nación. https://hdl.handle.net/20.500.12337/1181

Shin, D. y Shim, J. (2021). A systematic review on data mining for mathematics and science education [Una revisión sistemática sobre minería de datos para la educación en matemáticas y ciencias]. *International Journal of Science and Mathematics Education, 19*, 639-659. https://doi.org/10.1007/s10763-020-10085-7

Solís, M., Moreira, T., González, R., Fernández, T. y Hernández, M. (2018). Perspectives to predict dropout in university students with Machine Learning [Perspectivas para predecir la deserción escolar en estudiantes universitarios con Machine Learning]. *IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, 1-6. https://doi.org/10.1109/IWOBI.2018.8464191

Tinto, V. (1982). Definición de abandono escolar: una cuestión de perspectiva. *Nuevas Direcciones para la Investigación Institucional, 1982*(36), 3-15. https://doi.org/10.1002/ir.37019823603

Tinto, V. (1989). Definir la deserción: Una cuestión de perspectiva. *Revista de Educación Superior, 71*(18), 1-9. http://publicaciones.anuies.mx/pdfs/revista/Revista71_S1A3ES.pdf

Valencia, L. I., Guzmán, A. y Barragán, S. (2024). Deserción en programas de posgrado: un fenómeno poco explorado: una revisión exploratoria. *Educación Cogent, 11*(1), 1-20. https://doi.org/10.1080/2331186X.2024.2326705

Wang, W., Zhao, Y., Wu, Y. J. y Goh, M. (2023). Factores de abandono de los MOOCs: Una revisión bibliométrica. *Biblioteca Hi Tech, 41*(2), 432-453. https://doi.org/10.1108/LHT-06-2022-0306

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

Xu, C., Zhu, G., Ye, J. y Shu, J. (2022). Minería de datos educativos: Predicción de abandono en los MOOCs de XuetangX. *Cartas de Procesamiento Neuronal, 54*(4), 2885-2900. https://doi.org/10.1007/s11063-022-10745-5

Zamora-Araya, J. A. (2023a). *Modelo de un sistema de alerta temprana para reducir el abandono en el curso de Matemática General en la Universidad Nacional, Costa Rica* [Tesis de doctorado, Universidad Estatal a Distancia]. https://catalogosiidca.csuca.org/Record/UNED.000099829

Zamora-Araya, J. A. (2023b, noviembre). *Predicción del abandono temprano en estudiantes de nuevo ingreso en el curso de Matemática General utilizando algoritmos de aprendizaje supervisado* [Ponencia]. Congreso Latinoamericano sobre Abandono en Educación Superior (CLABES), Temuco, Chile. https://clabes.uct.cl/wp-content/uploads/2024/06/Acta-XII-CLABES-Revision-final.pdf

Zamora-Araya, J. A., Aguilar-Fernández, E. y Rodríguez-Pineda, M. (2023). ¿Cuándo el abandono universitario se convierte en exclusión educativa? *Revista Innovaciones Educativas, 25*(38), 97-115. http://dx.doi.org/10.22458/ie.v25i38.4212

Zamora-Araya, J. A., Gamboa, R., Hidalgo, R. y Castillo, M. (2020). Permanencia estudiantil en el curso de Matemática General de la Universidad Nacional, Costa Rica. *Actualidades Investigativas en Educación, 20*(1), 1-23. https://doi.org/10.15517/aie.v20i1.39815