ISBN: 978-628-96613-3-0. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2026.1.1.721

Predictive model based on machine learning for the early identification of student dropout in a public university

Pablo Aparicio-Montenegro[1]; Esparza Silva Milciades Roberto[1]; Huapaya Sotero Armando[1]
Maria Y. Garcia-Alvarez[2]; Andrea De La Cruz Garcia[2]

[1] Universidad Nacional Federico Villarreal, Perú, *paparicio@unfv.edu.pe, mesparza@unfv.edu.pe, ahuapaya@unfv.edu.pe*
[2]Universidad Tecnológica del Perú, Perú, *c19447@utp.edu.pe*, *u19313115@utp.edu.pe*

Abstract– *Student dropout represents a significant problem in public universities, with academic, economic, and social repercussions. This study aims to develop a predictive model based on machine learning techniques for the early identification of students at risk of dropping out. This will allow for the implementation of measures capable of reducing dropout rates through intervention strategies such as personalized tutoring and socioeconomic support. The data used includes academic, socioeconomic, and behavioral data corresponding to the periods 2015-2024. The methodology applied is quantitative and predictive, using algorithms linked to supervised learning techniques such as Logistic Regression, Decision Trees, Random Forest, and AdaBoost. The data were preprocessed through cleaning, Z-score normalization, and balancing with SMOTE. They were then divided into training (70%), validation (15%), and test (15%) sets, and k-fold cross-validation (k = 5) was applied. The performance of the models was evaluated using accuracy, precision, recall, F1-score, and area under the ROC curve. The results indicated that the AdaBoost model performed best, achieving an accuracy of approximately 89% and an AUC of 0.957, surpassing Random Forest (AUC = 0.902) and Logistic Regression (AUC = 0.948). Therefore, it is concluded that the proposed model could be considered a suitable tool for the early detection of students at risk of dropping out.*

Keywords-- Machine learning; student dropout; higher education; predictive model; public university.

Modelo predictivo basado en aprendizaje automático para la identificación temprana de la deserción estudiantil en una universidad pública

Pablo Aparicio-Montenegro[1]; Esparza Silva Milciades Roberto[1]; Huapaya Sotero Armando[1]
Maria Y. Garcia-Alvarez[2]; Andrea De La Cruz Garcia[2]

[1] Universidad Nacional Federico Villarreal, Perú, *paparicio@unfv.edu.pe, mesparza@unfv.edu.pe, ahuapaya@unfv.edu.pe*
[2]Universidad Tecnológica del Perú, Perú, *c19447@utp.edu.pe*, *u19313115@utp.edu.pe*

Resumen– *La deserción estudiantil representa un problema importante en las universidades públicas, con efectos académicos, económicos y sociales. Este estudio tiene como objetivo desarrollar un modelo predictivo basado en técnicas de machine learning para la identificación temprana de los estudiantes con riesgo de deserción, lo que permitirá la puesta en marcha de medidas capaces de reducir la deserción mediante estrategias de intervención como tutorías personalizadas y apoyo socioeconómico. Los datos utilizados académicos, socioeconómicos y conductuales que corresponden a los periodos 2015-2024. La metodología que se ha aplicado es de tipo cuantitativa y predictiva, utilizando algoritmos vinculados a la técnica de aprendizaje supervisado como Regresión Logística, Árboles de Decisión, Random Forest, y AdaBoost. Los datos fueron preprocesados mediante limpieza, la normalización Z-score, el balanceo con SMOTE y divididos en conjuntos de entrenamiento (70%), validación (15%) y prueba (15%), y se aplicó validación cruzada k-fold (k = 5). El desempeño de los modelos fueron evaluados mediante la exactitud, precisión, el recall, el F1-score y el área bajo la curva ROC. Los resultados obtenidos indicaron que el modelo AdaBoost presentó el mejor rendimiento, alcanzando una exactitud aproximada del 89% y un AUC de 0.957; superando a Random Forest (AUC = 0.902) y Regresión Logística (AUC = 0.948). De este modo, se concluye que el modelo propuesto podría ser considerado como una herramienta adecuada para la detección temprana de los estudiantes en riesgo de deserción.*

Palabras clave-- Aprendizaje automático; deserción estudiantil; educación superior; modelo predictivo; universidad pública.

I. Introducción

La deserción de los estudiantes en la educación superior es un grave problema que afecta en profundidad a las universidades, a los estudiantes y a la sociedad. En las universidades públicas de América Latina, las tasas de deserción varían entre el 20 % y el 50 %, según factores socioeconómicos, académicos e institucionales [1], [2]. Este hecho provoca una discontinuidad en el desarrollo académico de los estudiantes, y produce efectos negativos a nivel económico (desperdicio de recursos de los entes públicos) y a nivel social (poca movilidad social acompañada de escaso desarrollo humano). [3], [4] .

Desde una perspectiva académica, la deserción en la educación superior es ampliamente reconocida como un fenómeno multifactorial; diversos modelos teóricos, tanto clásicos como contemporáneos, han identificado factores académicos, sociales, psicológicos y económicos; entre los determinantes más característicos se erigen el bajo rendimiento académico, la mala integración social y académica durante los ciclos que se inician, la desmotivación, los problemas de salud mental y las dificultades económicas persistentes [5],[6],[7]. En el caso de las universidades públicas estas variables tienden a acentuarse con la presencia de estudiantes que provienen de cuartiles socioeconómicos vulnerables que deben atender a la vez las responsabilidades laborales, familiares y financieras que les llevan a aumentar el riesgo de desertar [8], [9].

Históricamente, las instituciones de educación superior han indagado el fenómeno de la deserción de los estudiantes aplicando metodologías de acercamientos descriptivos y reactivos, basados en indicadores históricos (por ejemplo, tasas de reprobación; promedio ponderados acumulado, etc.), estos métodos, permiten caracterizar y describir el fenómeno de forma retrospectiva, también poseen importantes limitaciones para su utilización con el fin de detectar a los estudiantes en riesgo de deserción y para intervenir correctamente. En este contexto es de suma importancia las técnicas predictivas y preventivas, fundamentadas en la analítica educativa y el crecimiento de los grandes volúmenes de datos institucional. [1], [10].

Los avances en el aprendizaje automático han cambiado significativamente la forma en que analizamos la información en el ámbito educativo. A diferencia de los métodos estadísticos tradicionales, estos algoritmos pueden manejar grandes volúmenes de datos diversos y complejos. Su

fortaleza está en descubrir patrones ocultos y relaciones no lineales entre múltiples factores, capturando la realidad de manera más sutil. Gracias a esta capacidad, ahora podemos hacer pronósticos más precisos y aplicables sobre el desempeño de los estudiantes. Incluso es posible identificar, con mayor anticipación y certeza, a aquellos que tienen un mayor riesgo de abandonar sus estudios, lo que abre la puerta a una ayuda más oportuna y personalizada.[9], [11]. Es por ello, los modelos predictivos basados en aprendizaje automático se han convertido en un apoyo clave para tomar decisiones en las universidades y para fortalecer las estrategias que buscan que los estudiantes sigan adelante con sus estudios. La utilidad de estos modelos se ve respaldada por estudios recientes. Un trabajo de investigación [12], por ejemplo, reveló que al analizar conjuntamente datos del perfil personal del estudiante, su historial académico y su interacción con las plataformas digitales de la universidad, se logra detectar con una precisión superior al 85% a aquellos en riesgo de dejar los estudios. Este hallazgo subraya una idea fundamental: el compromiso diario y la participación activa de los estudiantes son indicadores clave para anticipar y prevenir la deserción.

Además otro trabajo [13] demostró que estos sistemas no solo detectan a tiempo a los estudiantes con mayores dificultades, sino que también ayudan a diseñar intervenciones personalizadas y tempranas, lo que mejora significativamente las posibilidades de que continúen y culminen su formación.

A su vez, estudios como el de [14] la evidencia empírica sugiere que la disponibilidad de sistemas de información predictiva para universidades favorece la mejora de la priorización en las intervenciones y cómo esto se traduce en una mayor tasa de estudiantes que permanecen en la universidad. Un sistema de información predictivo se basa en algoritmos como Random Forest, AdaBoost o modelos bayesianos generalizados y permite estimar el nivel de riesgo de cada estudiante y activar alertas tempranas que ayudan a establecer la coordinación de las acciones de carácter académico, administrativo y de bienestar estudiantil. En esta misma línea, [15], subrayan la importancia de combinar modelos predictivos con enfoques explicativos, a fin de garantizar que los resultados generados por los algoritmos sean interpretables y útiles para la toma de decisiones estratégicas.

Sin embargo, y a pesar de los progresos de los que dan cuenta los estudios internacionales, continúa existiendo un hueco extenso en lo que a la aplicación sistemática de modelos predictivos de deserción en las universidades públicas latinoamericanas se refiere. Lo peculiar del contexto de la región, como por ejemplo la heterogeneidad del estudiantado, la escasez de financiación, la infraestructura tecnológica irregular o la escasa relación entre los sistemas de información académica, establece la necesidad de soluciones adaptadas a la realidad institucional y validadas con datos locales [16]. En ambos sentidos, se requieren modelos que, además de alcanzar un alto desempeño predictivo, sean operativamente viables, replicables y alineados con procesos de gestión académica actuales.

A pesar de la creciente oferta de investigaciones que profundizan en el uso de técnicas de aprendizaje automático para tratar de predecir la deserción de los estudiantes, la mayoría de los trabajos se centran en el trabajo en instituciones privadas, en la enseñanza virtual o en la utilización de bases de datos de carácter muy corto. Existe además una escasa evidencia empírica en el contexto de las universidades públicas de América Latina que utilice datos históricos largos y que valide la operatividad de los modelos predictivos como sistemas de alertas tempranas. En esta línea se responde la brecha de investigación que intenta abordar el desarrollo y la validación de un modelo predictivo robusto en función de las características distintivas de las universidades públicas de la región y sus peculiaridades socioeconómicas.

Con el fin de atender esta problemática, el presente trabajo plantea un modelo predictivo fundamentado en técnicas de aprendizaje automático y en la identificación precoz de la deserción de alumnos en una universidad pública, a partir de datos académicos, socioeconómicos y de datos del rendimiento histórico durante un periodo de 10 años. Su objetivo es prever la probabilidad de abandono en el estudiantado, permitiendo a la universidad poner en práctica medidas preventivas adecuadas, tales como tutorías personalizadas, programas de nivelación y apoyo socioeconómico dirigido, de tal forma que la investigación pretende reforzar la gestión de lo público fundamentado en los datos, además de aducir una prueba empírica que aporte evidencias sobre la aplicación de modelos de aprendizaje automático en entornos públicos de educación superior.

Por último, esta propuesta está dirigida tanto a investigadores como a gestores universitarios. Ofrece un modelo predictivo flexible, pensado para que pueda adaptarse y aplicarse en otras universidades públicas con contextos similares. Incorporar el aprendizaje automático en la gestión de una universidad no es solo una modernización tecnológica; es, sobre todo, una oportunidad estratégica. Permite actuar con mayor equidad, apoyando a quien más lo necesita; ayuda a optimizar los recursos disponibles; y, en definitiva, contribuye a construir una institución más justa y sostenible

II. METODOLOGÍA

A. Enfoque y diseño de la investigación

El presente estudio tiene un enfoque cuantitativo, de tipo aplicado y predictivo, cuyo objetivo es el desarrollo de un modelo de aprendizaje automático para la detección anticipada de la deserción estudiantil en una universidad pública. El diseño de la investigación adoptado es no experimental, retrospectivo y predictivo, basado en datos históricos institucionales, ya que las variables no se manipulan intencionadamente y se utilizan registros académicos de periodos anteriores [15]; [17].

El modelo utilizado es el adecuado para modelar patrones complejos vinculados con el abandono universitario y para calcular la probabilidad de la deserción antes de que esta suceda, como han evidenciado trabajos recientes de analítica educativa y minería de datos aplicada a la educación superior.[1], [10].

B. Contexto del estudio y población

El estudio se desarrolló en una universidad pública ubicada en Lima, Perú, utilizando registros académicos históricos correspondientes al periodo **2015–2024**. La población estuvo conformada por la totalidad de estudiantes matriculados en programas de pregrado durante dicho periodo. Debido a la disponibilidad completa de los registros, **no se aplicó muestreo**, considerándose la población total como conjunto de análisis, siguiendo prácticas recomendadas en estudios predictivos basados en datos institucionales.[11].

Las variables analizadas incluyeron indicadores académicos, conductuales y socioeconómicos, debidamente anonimizados para garantizar la confidencialidad de la información, en concordancia con principios éticos en investigación educativa [3].

C. Recolección y descripción de los datos

Los datos utilizados en esta investigación provienen de los sistemas académicos institucionales de una universidad pública de Lima, Perú, y la información recopilada abarcaba el histórico en el periodo 2015–2024. La elaboración de la base de datos incluyó registros de carácter académico, administrativo y socioeconómicos que eran producidos durante el trayecto universitario de los estudiantes, tomando un enfoque retrospectivo y con fundamentos en datos secundarios el cual es muy utilizado en la literatura de estudios de predicción del abandono estudiantil. [1].

El conjunto de datos estaba compuesto por variables de carácter académico (promedio ponderado acumulado, número de cursos desaprobados, condición académica), conductual (asistencia, continuidad de matrícula) y socioeconómico (condición laboral, nivel de estrés financiero) que por su parte la literatura menciona como aspectos significativos que pueden ser asociados al abandono universitario [2],[6].

Los datos fueron anonimizados y despersonalizados, eliminando datos personales que permitieran la identificación directa de los estudiantes, garantizando la confidencialidad y el uso ético de la información en concordancia con las recomendaciones internacionales de investigaciones educativas hechas en base a los datos institucionales [3].

La Fig. 1 explica el flujo de la investigación realizado para el estudio donde se presenta la recogida de datos, el pre-tratamiento de datos, el balanceo de clases, las particiones del conjunto de datos, el entrenamiento de los modelos (adicional) y la evaluación del desempeño.

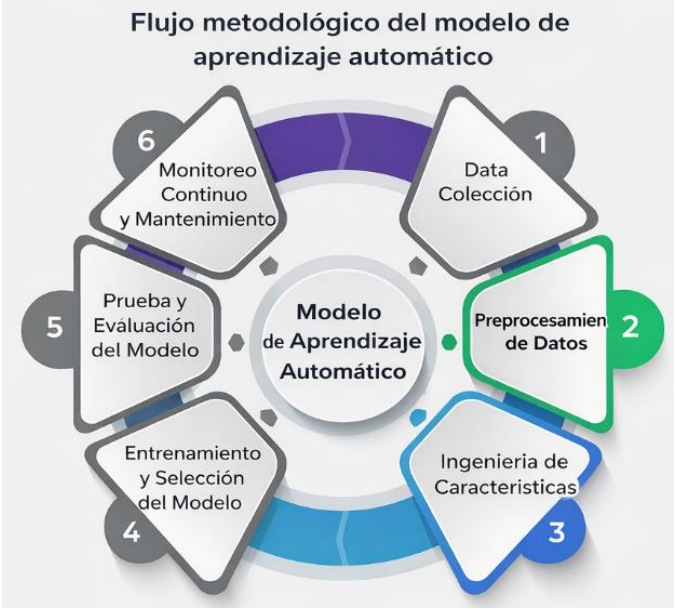

Fig. 1. Flujo metodológico del modelo predictivo utilizado en este estudio para la identificación temprana de estudiantes en riesgo de deserción.

D. Preprocesamiento de los datos

El preprocesamiento de datos incluyó la limpieza de registros inconsistentes, el tratamiento de valores faltantes y la normalización de variables numéricas mediante estandarización Z-score, con el fin de asegurar una contribución homogénea de las características al proceso de aprendizaje [18]; [19] .

Asimismo, se abordó el desbalance natural entre estudiantes desertores y no desertores mediante la técnica

SMOTE (Synthetic Minority Over-sampling Technique), ampliamente utilizada en problemas de clasificación desbalanceada ([20]; [21]).

La técnica SMOTE fue aplicada exclusivamente sobre el conjunto de entrenamiento, evitando la fuga de información hacia los conjuntos de validación y prueba, tal como recomiendan las buenas prácticas metodológicas en aprendizaje automático [17].

E. División del conjunto de datos y validación

Los datos fueron divididos en subconjuntos de entrenamiento (70 %), validación (15 %) y prueba (15 %). Durante la fase de entrenamiento se aplicó la validación cruzada k-fold (k=5) con la intención de mejorar la capacidad de generalización del modelo y reducir el riesgo de sobreajuste. [22]. El conjunto de prueba fue mantenido completamente independiente y no fue utilizado durante las etapas de selección de variables, balanceo de clases ni ajuste de hiperparámetros, garantizando así una evaluación objetiva e imparcial del desempeño predictivo final del modelo.

F. Construcción y selección del modelo predictivo

Se evaluaron diversos algoritmos de aprendizaje automático ampliamente utilizados en la literatura para la predicción de la deserción estudiantil, incluyendo Regresión Logística, Árboles de Decisión, Random Forest y AdaBoost, los cuales han demostrado un desempeño adecuado en problemas de clasificación binaria con datos académicos, conductuales y socioeconómicos [13], [14] . La selección de estos algoritmos permitió realizar un análisis comparativo considerando modelos interpretables y modelos basados en ensamblaje, con el fin de evaluar tanto el desempeño predictivo como la estabilidad de los resultados obtenidos.

Para la reducción de dimensionalidad, se aplicó un análisis de importancia de variables mediante el algoritmo Random Forest. A partir de un conjunto inicial de 165 variables, se seleccionaron las 25 características más relevantes. El número final de variables seleccionadas (n = 25) se determinó considerando el umbral acumulado de importancia de las características y la estabilidad del desempeño predictivo del modelo durante la validación cruzada, priorizando un equilibrio entre reducción de dimensionalidad, interpretabilidad y capacidad de generalización.

Los modelos fueron entrenados y ajustados exclusivamente sobre el conjunto de entrenamiento, mediante un proceso de optimización de hiperparámetros apoyado en validación cruzada k-fold (k = 5). Durante este proceso se priorizó la maximización del recall y del área bajo la curva ROC (AUC), considerando su relevancia en la detección temprana de estudiantes en riesgo, manteniendo al mismo tiempo un equilibrio adecuado con la precisión.

Si bien durante la fase exploratoria se consideraron modelos bayesianos generalizados, estos no fueron incluidos en el análisis comparativo final debido a un desempeño predictivo inferior y a una menor estabilidad frente a los modelos seleccionados. En consecuencia, el análisis final se concentró en aquellos algoritmos que demostraron un comportamiento más robusto y consistente.

Finalmente, la selección del modelo predictivo óptimo se realizó considerando de manera conjunta su desempeño cuantitativo, su estabilidad durante la validación cruzada y su viabilidad operativa para su eventual implementación como sistema de alerta temprana en una universidad pública. Adicionalmente, se evaluaron modelos de aprendizaje supervisado KNN y SVM con fines comparativos exploratorios, sin ser considerados en la selección final.

G. Evaluación del desempeño del modelo

Los resultados de los modelos fueron evaluados con métricas típicas en los problemas de clasificación binaria (exactitud -accuracy-, precisión, recall, F1-score y AUC -area under curve-), de forma que los indicadores ofrecían información relacionada tanto con la capacidad de discriminación de los modelos como con su aplicabilidad práctica para los casos en los que la detección de estudiantes en riesgo de deserción es una prioridad [12]; [15].

H. Implementación del modelo

Finalmente, el modelo elegido fue implantado en un entorno de prueba para simular su funcionamiento como un sistema de alerta temprana en el contexto institucional. Esto permitió evaluar la operatividad del modelo y su aplicabilidad para apoyar las estrategias de intervención académica, socioeconómica y tutorial para prevenir la deserción de los estudiantes [23]. La implantación se realizó en Python, utilizando distintas librerías consolidadas para el aprendizaje automático y el análisis de datos: Scikit-learn, Pandas, NumPy e Imbalanced-learn, en un entorno computacional local.

III. Resultados

A. Desempeño predictivo de los modelos

La evaluación del rendimiento de los modelos de machine learning tratados se realizó mediante métricas habituales de clasificación binaria (exactitud (accuracy), precisión, recall, F1 score y área bajo la curva ROC (AUC)), que son muy

usadas en trabajos de investigación que buscan predecir el abandono estudiantil en educación superior [12]; [15].

Los resultados que se muestran fueron obtenidos solamente del conjunto de prueba independiente, mientras que la validación cruzada k-fold (k = 5) se realizó sólo durante la fase de entrenamiento interna del modelo, para el ajuste de hiperparámetros, garantizando así una evaluación objetiva del desempeño predictivo final. La **Tabla 1** muestra el cuadro del desempeño comparativo de los algoritmos considerados: Regresión Logística, Árbol de Decisión, Random Forest y AdaBoost.

TABLA I.
DESEMPEÑO COMPARATIVO DE LOS MODELOS PREDICTIVOS

Modelo	Accuracy	Precision	Recall	F1-Score	AUC
Regresión Logística	0.88	0.86	0.84	0.85	0.948
Árbol de Decisión	0.85	0.83	0.81	0.82	0.881
Random Forest	0.92	0.91	0.89	0.90	0.902
AdaBoost	0.89	0.87	0.85	0.86	0.957
KNN	0.78	0.80	0.74	0.76	0.746
SVM	0.75	0.78	0.72	0.75	0.696

En este estudio, se escoge el modelo óptimo en función de un criterio que es el correcto en relación con el objetivo del problema a tratar, se priorizan la capacidad de detección de los estudiantes en riesgo (recall) y la capacidad discriminativa general del modelo (AUC), más que métricas como la exactitud. Justificamos esta elección debido a que en el contexto educativo, el costo de no identificar a un estudiante en riesgo (falso negativo) es significativamente mayor que clasificar incorrectamente a un estudiante sin riesgo.

Siguiendo este criterio, el modelo de AdaBoost es el que presenta mejor desempeño global debido al hecho de que presenta el mayor valor de AUC (0.957) además de tener un equilibrio correcto entre la precisión (0.87) y el recall (0.85), lo que evidencia una alta capacidad para discriminar correctamente estudiantes en riesgo y no en riesgo.

Por su parte, el modelo de Random Forest obtuvo mayor exactitud (0.92) y también destaca en métricas tradicionales como la precisión y el F1-score, sin embargo, su AUC (0.902) es inferior a la de AdaBoost, lo que se puede interpretar como una menor capacidad discriminativa global del modelo respecto al anterior.

En consecuencia, el modelo AdaBoost fue seleccionado como el modelo óptimo desde el punto de vista del desempeño predictivo, al presentar la mayor capacidad discriminativa (AUC = 0.957) y un adecuado equilibrio entre métricas.

Sin embargo, el modelo de Random Forest se utilizó de manera complementaria para obtener la importancia de las variables y explicar el comportamiento del modelo, pues proporciona una mayor estabilidad, robustez y facilidad de la interpretación, aspectos especialmente relevantes en contextos institucionales, donde la explicabilidad resulta fundamental para la toma de decisiones.

B. Comparación entre algoritmos de aprendizaje automatico

La comparación realizada para los modelos de evaluación presentados en el presente trabajo muestra diferencias considerables en su capacidad predictiva y estabilidad . En este sentido , el modelo AdaBoost es el que obtiene mejor rendimiento en cuanto a discriminación global (AUC = 0.957), en línea con el criterio de selección adoptado y que persigue , en última instancia , maximizar la identificación de los estudiantes en riesgo de no continuar sus estudios.

El modelo Random Forest , en cambio , muestra un rendimiento altamente competitivo, sobre todo en las métricas exactitud, precisión y F1 , lo que concuerda con la literatura especializada, donde se pone de manifiesto que los métodos basados en combinaciones de árboles son muy robustos para detectar relaciones no lineales e interacciones entre las variables académicas, de conducta y socio - económicas [13], [14]. En este sentido , los ensamblajes de árboles de decisión contribuyen a reducir el sesgo y la varianza, lo que mejora la capacidad de su generalización.

La Regresión Logística, aunque presenta valores ligeramente inferiores a los anteriores , en cuanto a exactitud y AUC respecto a los modelos de ensamblaje, muestra un equilibrio adecuado entre la precisión y recall, lo que confirma su relevancia como modelo interpretable en aquellos escenarios donde la explicabilidad sea un prerequisito indispensable para la toma de decisiones [15]. En contraste , el modelo de Árbol de Decisión presenta sobreajuste, que se manifiesta en una menor estabilidad de su rendimiento predictivo.

C. Importancia de variables predictoras

El análisis de relevancia de las variables obtenido mediante el modelo Random Forest permite determinar las variables que gozan de una mayor incidencia para predecir la deserción estudiantil. Entre las que gozan de una mayor relevancia predictiva se encuentran el promedio ponderado acumulativo (CGPA), el número de cursos desaprobados, el índice de riesgo académico, la tasa de asistencia, y el nivel de

estrés financiero. Los resultados obtenidos son concordantes con la existencia de evidencia tanto teórica como empírica que pone de manifiesto la relevancia determinante que para la permanencia o no de los estudiantes en educación superior tienen el rendimiento académico y las condiciones socioeconómicas [2], [6], [16] . En particular, el rendimiento académico bajo, así como la cantidad de asignaturas desaprobadas, son considerados signos tempranos relevantes del riesgo de abandono, en medio de los elementos socioeconómicos que operan como indicadores estructurales que tensionan este riesgo en el contexto específico de universidades públicas.

La identificación de estas variables clave resulta fundamental para dar forma a las estrategias institucionales de intervención temprana desde el ámbito de la tutoría académica, la atención personalizada y la atención socioeconómica de enfoque. Asimismo, la matriz de correlación de la Figura 2 refleja la existencia de importantes relaciones entre las variables que poseen un mayor valor predictivo respecto a la deserción estudiantil. Variables como el promedio ponderado acumulado (CGPA), la cantidad de cursos desaprobados y el índice de estrés financiero presentan correlaciones considerables respecto a la variable de deserción. Esta información es importante para el entendimiento de qué factores son los que más impactan al comportamiento de deserción y cómo se correlacionan entre ellos. De esta manera, contribuye a realizar una mejor selección de las variables a tener en cuenta a ser utilizadas en los modelos predictivos.

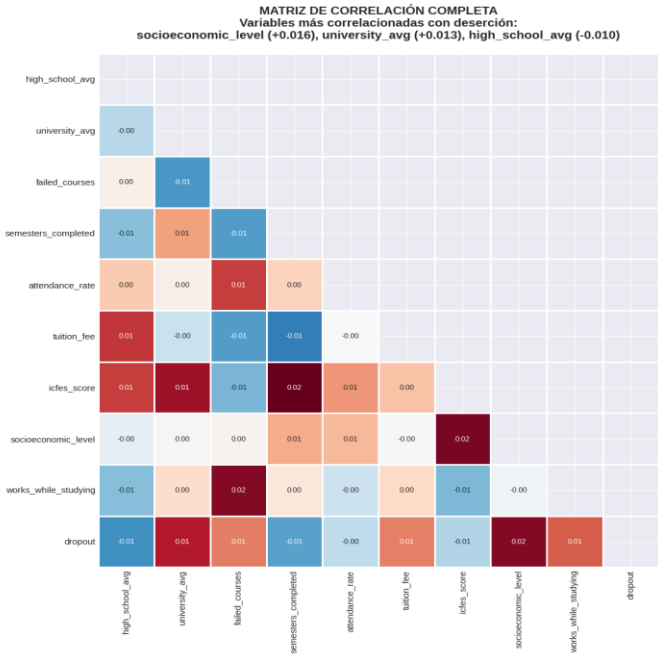

Fig. 2. Matriz de Correlación

D. Predicción de casos individuales

Con el objetivo de llevar a cabo un análisis sobre la capacidad de generalización del modelo preferido, se llevaron a cabo pruebas usando registros de alumnos que no habían sido utilizados para el aprendizaje. Como resultado, el método Random Forest consiguió clasificar correctamente los alumnos que tenían una alta probabilidad de abandono así como aquellos con baja probabilidad de abandono, con predicciones coherentes con sus características académicas y conductuales.

Esta capacidad para ofrecer pronósticos individuales fiables permitía dotar del potencial operativo al modelo como sistema de alerta temprana, gracias a la posibilidad de detectar de forma anticipada a los alumnos en riesgo y poder tomar decisiones desde la institución para ayudar a mejorar la retención estudiantil. Los resultados obtenidos son coincidentes con estudios recientes que destacan la efectividad de las predictivas basadas en el aprendizaje automático, siempre que fuesen integradas en el sistema de información académica de forma adecuada.[12], [23].

La Figura 3 muestra la curva ROC para cada uno de los modelos evaluados. Los modelos de Regresión Logística (AUC = 0.948), Árbol de Decisión (AUC = 0.881), Random Forest (AUC = 0.902), AdaBoost (AUC = 0.957), K-Vecinos (AUC = 0.746), SVM (AUC = 0.696), y Clasificador Aleatorio (AUC = 0.500) se evaluaron para determinar su capacidad discriminativa. El modelo AdaBoost resultó ser el más efectivo con un AUC de 0.957, seguido de la Regresión Logística con un AUC de 0.948, mientras que el modelo SVM presentó el peor desempeño con un AUC de 0.696.

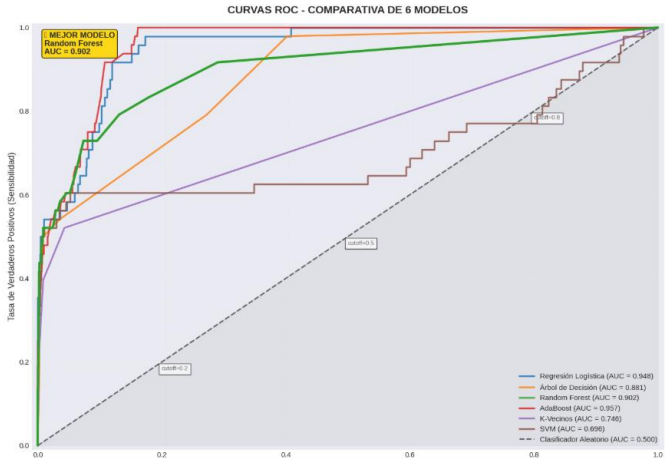

Fig. 3. Curva ROC según modelo

Síntesis de resultados

En síntesis, la información de los resultados demuestra que el modelo de Machine Learning propuesto en este documento tiene una buena capacidad para detectar estudiantes en riesgo de abandono. La mejor eficacia del algoritmo **AdaBoost**, junto con la estabilidad observada en los modelos evaluados y la relevancia de las variables predictoras identificadas, respalda la viabilidad de implementar este enfoque en universidades públicas como sistema de alerta temprana.

IV. Discusión

Los resultados de este estudio corroboran la validez y eficiencia de los modelos de aprendizaje automático para la predicción de la deserción de los estudiantes universitarios en universidades públicas. La alta capacidad predictiva de los modelos estudiados, y en particular del modelo AdaBoost, seguido de Random Forest y la Regresión Logística, indica que los modelos predictivos con información académica histórica, así como la información socioeconómica y conductual, permiten obtener predicciones en la probabilidad de abandono universitario muy superiores a las limitaciones de los modelos descriptivos tradicionales empleados en las universidades.

Los resultados del modelo AdaBoost, que presentó la mayor capacidad discriminativa global (AUC), confirman lo que se ha reportado en estudios anteriores, que han evidenciado que los algoritmos de ensamblaje basados en árboles son robustos en la clasificación de contextos educativos complejos [13],[14]. Tal comportamiento es el resultado de la habilidad del algoritmo para modelar relaciones no lineales e interacciones complejas de las variables explicativas, aunque también es un aspecto de gran relevancia en un fenómeno multifactorial como la deserción de los estudiantes. Asimismo, el buen equilibrio entre precisión y recall observado en la Regresión Logística reafirma su utilidad como modelo interpretable, facilitando la comprensión de los factores asociados al abandono académico, tal como señalan [15].

Estos resultados evidencian que el desempeño óptimo de los modelos depende del criterio de evaluación adoptado, siendo el AUC una métrica particularmente relevante en contextos de clasificación desbalanceada y detección temprana.

La evaluación de la importancia de las variables ha dejado ver que los índices académicos, como el promedio ponderado acumulado, el historial de cursos desaprobados y el indicador de riesgo académico son los principales predictores de la deserción de los alumnos, lo cual se articula con los teóricos clásicos de la deserción en estudiantes universitarios que dan, ya lo dicho, un lugar privilegiado al rendimiento académico y a la integración académica temprana respecto de la permanencia de los estudiantes [6],[7]. De un modo complementario, la relevancia del índice de estrés financiero y de las variables socioeconómicas avala que el abandono de estudios en universidades públicas de Latinoamérica no puede ser estudiado solamente desde la perspectiva académica sino que se debe analizar como un fenómeno estructural vinculado a condiciones de vulnerabilidad social, tal como demuestran investigaciones recientes sobre la región [2], [16].

Desde una perspectiva metodológica, la implementación de técnicas de preprocesamiento de datos como la normalización de las variables y el balanceo de clases a partir del SMOTE ha sido fundamental para poder mejorar la capacidad predictiva de los modelos, ya que la literatura es bastante contundente sobre la importancia de tratar el desbalance en clases en problemas de predicción de abandono, debido a que la menor proporción de estudiantes desertores puede inducir sesgos en el entrenamiento de los algoritmos si no se implementan estrategias. [20],[21]. De este modo, los resultados de esta investigación destacan la relevancia de una fase de preparación de los datos rigurosa como condición necesaria para la generación de modelos confiables y generalizables.

.

Un elemento a tener en cuenta es la idoneidad de usar el modelo construido como un sistema de alerta temprana en la gestión universitaria; la correcta clasificación de los casos individuales que no formaban parte del proceso de entrenamiento justifica su capacidad de generalización a estudiantes que no formaron parte del proceso de clasificación de casos, constituyendo así una alternativa a seguir para la toma de decisiones institucionales. Este resultado es coherente con los articulos de investigación de [1], [12] afirman que la verdadera utilidad de los modelos predictivos se encuentra en su incorporación efectiva en los sistemas de información académica, así como en su combinación con tácticas de intervención a tiempo, como programas de tutoría académica, asistencia socioeconómica específica y programas de acompañamiento.

Los resultados del modelo predictivo de IA son muy positivos y constituyen un elemento importante tanto en el ámbito académico como a lo institucional —apoyado en pruebas empíricas que lo avalan como herramienta para propiciar un aumento en la permanencia de estudiantes de universidades públicas, en consonancia con las corrientes actuales de gestión de datos educativos—. Aunque las evidencias sugieren que su análisis tiene que ser ajustado al contexto universitario particular, ya sea de forma institucional, como en los diseños curriculares y la realidad socioeconómica del alumnado. Para que pueda ser implementado en otras universidades es necesario hacer primero la validación y los

ajustes necesarios con datos propios, en combinación con otras estrategias de inclusión y soporte inicial, tal como advierte sobre la necesidad de ajustar a contextos culturales e institucionales.

V. Conclusiones

La presente investigación propone un modelo basado en aprendizaje automático para identificar de manera anticipada a los estudiantes de universidades públicas que podrían estar en riesgo potencial de abandonar sus estudios. Para ello, se emplearon datos académica, socioeconómicos y de conducta de los estudiantes.

Los hallazgos indican que el modelo AdaBoost fue el que ofreció el mejor desempeño predictivo, alcanzando una precisión muy alta (AUC = 0.957) y una capacidad destacada para identificar correctamente a los estudiantes en riesgo. Sin embargo, el modelo Random Forest también demostró una excelente capacidad predictiva (AUC = 0.902), lo que lo convierte en una opción muy válida y práctica para instituciones que cuenten con recursos informáticos más limitados.

Estos hallazgos permiten sustentar la implementación de sistemas de alerta temprana que contribuyan a las universidades a dirigir esfuerzos de apoyo académico y socioeconómico de manera más eficiente y personalizada. De esta forma, no solo se optimizan los recursos institucionales, sino que se contribuye directamente a mejorar la retención de los estudiantes. En resumidas cuentas, este tipo de instrumentos ofrece a los gestores educativos información valiosa y pertinente para fundamentar sus decisiones y actuar a tiempo. Los resultados, en suma, demuestran que los modelos de machine learning constituyen una opción válida para el diseño de sistemas de alerta temprana en universidades públicas, promoviendo así una gestión académica basada en la exploración sistemática de datos.

Limitaciones y trabajos futuros

Este estudio tiene como limitación, el estar basado en los datos de una sola universidad, sus conclusiones podrían no aplicarse directamente a otras instituciones con contextos diferentes, el modelo actual no considera aspectos personales, como la motivación, el estrés, el estado anímico de los estudiantes, factores que influyen en la decisión de los estudiantes en desertar.

Como líneas futuras de investigación, se propone validar el modelo en otras universidades públicas para asegurar su utilidad, y enriquecerlo con técnicas más avanzadas, como el aprendizaje profundo. El objetivo final es lograr un modelo no solo predictivo, sino también aplicable para los gestores universitarios entender el "por qué" detrás de las predicciones y tomar decisiones certeras y transparentes.

Referencias

[1] C. F. De Oliveira, S. R. Sobral, M. J. Ferreira, y F. Moreira, «How Does Learning Analytics Contribute to Prevent Students' Dropout in Higher Education: A Systematic Literature Review», *Big Data Cogn. Comput.*, vol. 5, n.º 4, p. 64, nov. 2021, doi: 10.3390/bdcc5040064.

[2] B. R. Villegas y L. A. Núñez Lira, «Factores asociados a la deserción estudiantil en el ámbito universitario. Una revisión sistemática 2018-2023», *RIDE Rev. Iberoam. Para Investig. El Desarro. Educ.*, vol. 14, n.º 28, may 2024, doi: 10.23913/ride.v14i28.1923.

[3] *Global Education Monitoring Report 2019: Migration, Displacement and Education – Building Bridges, not Walls*. UNESCO, 2019. doi: 10.54676/XDZD4287.

[4] J. A. Castro-Martínez y G. Machuca-Téllez, «La deserción universitaria en América Latina: una perspectiva ecológica», *Estud. Pedagógicos Valdivia*, vol. 49, n.º 2, pp. 87-108, 2023, doi: 10.4067/S0718-07052023000200087.

[5] N. Reyes-González y A. L. Meneses-Báez, «Psychosocial Factors Associated With Dropout, Performance and Adjustment in University First Year: A Systematic Review of Reviews», *Rev. Electrónica Educ.*, vol. 28, n.º 2, pp. 1-23, ago. 2024, doi: 10.15359/ree.28-2.18435.

[6] R. Q. Apumayta, J. C. Cayllahua, A. C. Pari, V. I. Choque, J. C. C. Valverde, y D. H. Ataypoma, «[version 2; peer review: 2 approved] Previously titled: University Dropout: A Systematic Review of the Main Determinant Factors», *F1000Research*, vol. 13, 2024, doi: https://doi.org/10.12688/f1000research.154263.2.

[7] J. G. Piepenburg y J. Beckmann, «The relevance of social and academic integration for students' dropout decisions. Evidence from a factorial survey in Germany», *Eur. J. High. Educ.*, vol. 12, n.º 3, pp. 255-276, jul. 2022, doi: 10.1080/21568235.2021.1930089.

[8] T. Stylianou y A. Milidis, «The socioeconomic determinants of University dropouts: The case of Greece», *J. Infrastruct. Policy Dev.*, vol. 8, n.º 6, p. 3729, jun. 2024, doi: 10.24294/jipd.v8i6.3729.

[9] A. Valencia-Arias, S. Chalela, M. Cadavid-Orrego, A. Gallegos, M. Benjumea-Arias, y D. Y. Rodríguez-Salazar, «University Dropout Model for Developing Countries: A Colombian Context Approach», *Behav. Sci.*, vol. 13, n.º 5, p. 382, may 2023, doi: 10.3390/bs13050382.

[10] C. Márquez-Vera, A. Cano, C. Romero, y S. Ventura, «Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data», *Appl. Intell.*, vol. 38, n.º 3, pp. 315-330, abr. 2013, doi: 10.1007/s10489-012-0374-8.

[11] D. E. Moreira Da Silva, E. J. Solteiro Pires, A. Reis, P. B. De Moura Oliveira, y J. Barroso, «Forecasting Students Dropout: A UTAD pu Study», *Future Internet*, vol. 14, n.º 3, p. 76, feb. 2022, doi: 10.3390/fi14030076.

[12] S. C. Matz, C. S. Bukow, H. Peters, C. Deacons, A. Dinu, y C. Stachl, «Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics», *Sci. Rep.*, vol. 13, n.º 1, p. 5705, abr. 2023, doi: 10.1038/s41598-023-32484-w.

[13] C. L. Kok, C. K. Ho, L. Chen, Y. Y. Koh, y B. Tian, «A Novel Predictive Modeling for Student Attrition Utilizing Machine Learning and Sustainable Big Data Analytics», *Appl. Sci.*, vol. 14, n.º 21, p. 9633, oct. 2024, doi: 10.3390/app14219633.

[14] S. Guzmán-Castillo *et al.*, «Implementation of a Predictive Information System for University Dropout Prevention», *Procedia Comput. Sci.*, vol. 198, pp. 566-571, 2022, doi: 10.1016/j.procs.2021.12.287.

[15] Delen D., Davazdahemami B., y Dezfouli E., «Predicting and Mitigating Freshmen Student Attrition: A Local-Explainable Machine Learning Framework», *Inf Syst Front 26*, vol. 26, pp. 641-662, 2024, doi: https://doi.org/10.1007/s10796-023-10397-3.
[16] A. M. Rahmani, W. Groot, y H. Rahmani, «Dropout in online higher education: a systematic literature review», *Int. J. Educ. Technol. High. Educ.*, vol. 21, n.º 1, p. 19, mar. 2024, doi: 10.1186/s41239-024-00450-9.
[17] J.-E. Aspee, M. Elias, y J.-A. González-Campos, «Predicción de la deserción académica mediante aprendizaje automático supervisado: el caso de una universidad chilena», *Estud. Sobre Educ.*, nov. 2025, doi: 10.15581/004.51.004.
[18] S. García, J. Luengo, y F. Herrera, *Data Preprocessing in Data Mining*, vol. 72. en Intelligent Systems Reference Library, vol. 72. Cham: Springer International Publishing, 2015. doi: 10.1007/978-3-319-10247-4.
[19] Mariuxi Guillen-Intriago y Alcivar-Ceballos, Roberth, «Impact of data normalization on the accuracy of supervised learning models», *RIEMAT*, vol. 10, n.º 2, pp. 59-79, 2025, doi: https://doi.org/10.33936/riemat.v.10i2.7853.
[20] N. V. Chawla, K. W. Bowyer, L. O. Hall, y W. P. Kegelmeyer, «SMOTE: Synthetic Minority Over-sampling Technique», *J. Artif. Intell. Res.*, vol. 16, pp. 321-357, jun. 2002, doi: 10.1613/jair.953.
[21] E. F. Swana, W. Doorsamy, y P. Bokoro, «Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset», *Sensors*, vol. 22, n.º 9, p. 3246, abr. 2022, doi: 10.3390/s22093246.
[22] L. A. Yates, Z. Aandahl, S. A. Richards, y B. W. Brook, «Cross validation for model selection: A review with examples from ecology», *Ecol. Monogr.*, vol. 93, n.º 1, p. e1557, feb. 2023, doi: 10.1002/ecm.1557.
[23] N. J. Dia *et al.*, «EduGuard RetainX: An advanced analytical dashboard for predicting and improving student retention in tertiary education», *SoftwareX*, vol. 29, p. 102057, feb. 2025, doi: 10.1016/j.softx.2025.102057.