Revista Ciencia, Tecnología e Innovación en Ingeniería y Arquitectura
2(1) (2026), pp. 19-29
Facultad de Ingeniería y Arquitectura
Universidad de El Salvador
https://revistas.ues.edu.sv/index.php/cti-fia

CTI-FIA

Artículo | Article

Optimización de la predicción de retención estudiantil en la educación superior: un análisis comparativo de modelos de aprendizaje automático supervisados y no supervisado

Optimizing student retention prediction in higher education: A comparative analysis of supervised and unsupervised machine learning models

Franklin Ivan Argueta Bermudez[1,2,], Wilber Alexander Ortiz Cortez[1], Miguel García Torres[1]

1 Facultad de Ingeniería y Arquitectura, Universidad de El Salvador
2 Escuela Politécnica Superior, España, Universidad Pablo de Olavide
3 https://orcid.org/0000-0002-7463-9873

Correspondencia: ***fivan.bermudez@gmail.com***
Enviado: ***20 de mayo de 2026*** Aceptado: ***30 de mayo de 2026***

DOI: https://doi.org/10.66778/CTI.v02n01.03

PALABRAS CLAVE

Retención estudiantil, deserción, Machine Learning, educación superior, modelos predictivos.

RESUMEN

Esta investigación aborda el problema de la deserción estudiantil en la educación superior en El Salvador, utilizando modelos de aprendizaje automático. Para ello, se empleó un conjunto de datos históricos proveniente de una universidad privada, que incluye más de 83,000 registros de estudiantes entre los años 2019 y 2024. En la fase de análisis exploratorio, se aplicaron técnicas no supervisadas como Análisis de Componentes Principales (PCA), K-means y agrupamiento jerárquico, con el objetivo de segmentar a los estudiantes e identificar perfiles con características demográficas, académicas y económicas similares. Posteriormente, se implementaron modelos predictivos supervisados, entre ellos SVM, KNN, Random Forest, AdaBoost y XGBoost, evaluando su rendimiento mediante métricas como precisión y AUC-ROC. Los resultados indicaron que Random Forest ofrece una mayor capacidad predictiva, mientras que las técnicas de clusterización permitieron identificar grupos de riesgo sin necesidad de etiquetas previas. Se observó que los estudiantes en modalidad presencial, con bajo rendimiento académico y autofinanciamiento, presentan una mayor probabilidad de abandono. La investigación concluye que la combinación de análisis exploratorio y modelos predictivos facilita el diseño de estrategias efectivas para mejorar la retención estudiantil. Además, se destaca el valor del uso institucional de estas herramientas para desarrollar intervenciones proactivas, especialmente dirigidas a estudiantes de primer año o con condiciones académicas y económicas vulnerables.

KEYWORDS

Student retention, dropout rate, machine learning, higher education, predictive models.

ABSTRACT

This research addresses the problem of student dropout in higher education in El Salvador, using machine learning models. To do so, a historical dataset from a private university was used, which includes more than 83,000 student records between the years 2019 and 2024. In the exploratory analysis phase, unsupervised techniques such as Principal Component Analysis (PCA), K-means, and hierarchical clustering were applied to segment students and identify profiles with similar demographic, academic, and economic characteristics. Subsequently, supervised predictive models were implemented, including SVM, KNN, Random Forest, AdaBoost, and XGBoost, evaluating their performance using metrics such as accuracy and AUC-ROC.

The results indicated that Random Forest offers greater predictive capacity, while clustering techniques allowed identifying risk groups without the need for prior labels. It was observed that students enrolled in in-person classes, with low academic performance and self-funded programs, are more likely to drop out. The research concludes that the combination of exploratory analysis and predictive models facilitates the design of effective strategies to improve student retention. Furthermore, it highlights the value of institutional use of these tools to develop proactive interventions, especially those targeting first-year students or those with vulnerable academic and economic circumstances.

Publicado por la Facultad de Ingeniería y Arquitectura de la Universidad de El Salvador.
Este es un artículo *Open Access* bajo la licencia *Creative Commons* Atribución 4.0 Internacional (CC BY 4.0)

INTRODUCCIÓN

La educación superior desempeña un papel estratégico en el desarrollo económico, social y cultural de los países. Factores como el acceso, la permanencia y la finalización oportuna de los estudios universitarios son fundamentales para formar profesionales que contribuyan al progreso de sus comunidades. No obstante, la deserción estudiantil continúa siendo uno de los desafíos más persistentes en las instituciones de educación superior, generando impactos negativos a nivel individual, institucional y social. En el caso de El Salvador, este fenómeno se ha ido normalizando debido a las desigualdades socioeconómicas, la limitada disponibilidad de recursos educativos y en su momento las secuelas dejadas por la pandemia de COVID-19. Estas condiciones afectan especialmente a estudiantes de contextos vulnerables, reduciendo sus posibilidades de permanecer y culminar con éxito su formación universitaria.

Diversas investigaciones han documentado múltiples causas de la deserción, entre ellas el bajo rendimiento académico, las dificultades económicas, los problemas familiares y la falta de alineación entre las expectativas del estudiante y la experiencia universitaria. Aunque muchas instituciones han implementado programas de tutoría, orientación y apoyo financiero, estas iniciativas suelen carecer de mecanismos efectivos para identificar con anticipación a los estudiantes en mayor riesgo de abandono. La creciente digitalización de los procesos académicos y administrativos ha generado una abundante disponibilidad de datos estudiantiles. Este nuevo panorama ofrece una oportunidad valiosa para aplicar herramientas de análisis de datos y aprendizaje automático que permitan detectar patrones complejos y anticipar comportamientos como la deserción. A través de estos enfoques, es posible desarrollar sistemas de alerta temprana que faciliten intervenciones proactivas y personalizadas.

En este contexto, la presente investigación tiene como objetivo general optimizar la predicción de la retención estudiantil mediante un análisis comparativo de modelos de aprendizaje automático supervisados y no supervisados, aplicados a un conjunto de datos históricos proporcionado por una universidad privada en El Salvador, que comprende más de 83,000 registros entre 2019 y 2024. El estudio se aplica desde un enfoque cuantitativo y exploratorio, a través de técnicas de segmentación y modelado predictivo orientadas a identificar perfiles estudiantiles con características similares de riesgo y predecir la deserción. Esta perspectiva permite generar evidencia útil para la toma de decisiones estratégicas en contextos institucionales con recursos limitados.

Además del aporte metodológico, esta investigación busca fomentar una cultura institucional basada en el uso sistemático de los datos para la mejora continua. En lugar de adoptar respuestas reactivas, se plantea un modelo de gestión proactiva que reconozca la diversidad del estudiantado y promueva intervenciones oportunas, especialmente dirigidas a estudiantes del primer ciclo o con condiciones académicas y económicas desfavorables. A nivel internacional, diversas universidades han demostrado que la integración de modelos predictivos en sus sistemas de gestión contribuye significativamente a mejorar los índices de retención y titulación. En sintonía con estas experiencias, el presente estudio se alinea con la tendencia global hacia una educación superior más analítica, digital y centrada en el estudiante, en la que los datos no solo documentan el pasado, sino que también proyectan el futuro educativo.

METODOLOGÍA

A. Estudios Similares

En el ámbito comercial, el concepto de valor del cliente se refiere a la cantidad de ingresos que un cliente genera para una empresa durante el período en que mantiene una relación activa con ella. Este valor no solo está vinculado a la rentabilidad inmediata, sino también al impacto que dicho cliente puede tener sobre la estabilidad y el crecimiento a largo plazo de la organización. Según Kotler y Keller (2016), comprender y gestionar el valor del cliente es esencial para que las empresas orienten sus estrategias hacia la retención de los segmentos más rentables, optimizando así el uso de recursos y maximizando beneficios.

En el contexto educativo, un concepto análogo es el valor del estudiante, entendido como la contribución que la permanencia de un estudiante en una institución tiene sobre sus ingresos, sostenibilidad financiera y reputación académica (Fahd et al., 2022). Este paralelismo entre ambos ámbitos pone de manifiesto que, aunque el producto o servicio sea distinto, la retención sigue siendo un factor crítico para la viabilidad organizacional. Sin embargo, existen diferentes factores que influyen en la permanencia. Mientras que en el sector comercial la retención depende principalmente de la calidad del producto o servicio, la experiencia del cliente y las políticas de fidelización (Kotler & Armstrong, 2018), en la educación superior la decisión de continuar o abandonar los estudios se ve influenciada por diferentes variables de tipo académicas, socioeconómicas, demográficas y psicosociales

(Tinto, 2012).

En este sentido, diversos autores han identificado y clasificado los factores que inciden en la retención estudiantil. Los factores demográficos incluyen la edad, el género, el nivel socioeconómico y la situación laboral del estudiante. Los factores académicos se relacionan con el rendimiento previo, la satisfacción con el programa de estudios y la percepción de la carga académica. Por último, los factores psicosociales abarcan aspectos como la motivación personal, la integración social dentro de la comunidad universitaria y el apoyo familiar (Astin, 1999; Tinto, 2012). Cabrera et al. (2006) señalan que la probabilidad de abandono aumenta significativamente cuando los estudiantes enfrentan una combinación de riesgos, como bajos ingresos familiares y dificultades académicas, lo que resalta la necesidad de una detección temprana de dichos factores.

Para abordar este reto, la literatura ha documentado el uso de modelos estadísticos y de aprendizaje automático que permiten identificar patrones y predecir la retención. Entre los enfoques más tradicionales, la regresión logística ha sido ampliamente utilizada para estimar la probabilidad de abandono en función de un conjunto de variables independientes, ofreciendo interpretaciones claras de la influencia de cada factor (Cortez & Silva, 2008). En paralelo, las técnicas de aprendizaje no supervisado, como el algoritmo K-means, han demostrado utilidad para agrupar estudiantes según similitudes en sus características, revelando patrones que no siempre son evidentes mediante análisis univariados o bivariados (Kurniawan & Taufiq, 2019).

Estudios comparativos han evaluado el rendimiento de diferentes modelos predictivos. Vandamme et al. (2017) confrontaron modelos supervisados como la regresión logística y los árboles de decisión, encontrando que estos últimos ofrecían una mayor precisión en la identificación de estudiantes en riesgo de deserción. Kurniawan y Taufiq (2019), por su parte, aplicaron técnicas de clustering para segmentar poblaciones estudiantiles y lograron identificar grupos con alto riesgo sin necesidad de etiquetas previas. Más recientemente, Arqawi et al. (2022) exploraron un enfoque híbrido que combina aprendizaje automático tradicional con modelos de aprendizaje profundo, alcanzando altos niveles de precisión y reduciendo la tasa de falsos positivos, lo que se traduce en intervenciones más focalizadas y efectivas.

Estos hallazgos reflejan que el uso de herramientas analíticas avanzadas no solo aporta valor académico, sino que también tiene implicaciones prácticas para la gestión institucional. La detección temprana de estudiantes en riesgo, seguida de intervenciones personalizadas como tutorías académicas, orientación vocacional o apoyos económicos puede contribuir de manera significativa a mejorar las tasas de retención (Martínez et al., 2021). En este contexto, la presente investigación se alinea con estos esfuerzos, adoptando un enfoque cuantitativo y un diseño mixto que combina análisis descriptivo y exploratorio con modelado predictivo. El objetivo es doble: por un lado, caracterizar los patrones de deserción estudiantil, y por otro, estimar la retención a partir de variables sociodemográficas, académicas y económicas. Para ello, se emplean técnicas no supervisadas, que permiten segmentar la población en grupos con características similares, y supervisadas, que facilitan la predicción precisa del abandono estudiantil.

B. Datos y fuentes

El conjunto de datos utilizado en esta investigación fue extraído de los registros académicos de una universidad privada en El Salvador, abarcando el periodo 2019-2024. Este dataset, compuesto por 83,175 registros y 31 variables, constituye la base fundamental para el análisis y la predicción de la retención estudiantil.

Las variables se agrupan en cuatro categorías principales:

- Sociodemográficas: género, edad, departamento de residencia, estado familiar.
- Académicas: facultad, carrera, nivel académico, rendimiento (CUM), modalidad, asignaturas aprobadas y reprobadas.
- Económicas: tipo de financiamiento.
- Históricas: año y semestre de ingreso, tipo de ingreso, continuidad académica.

Todos los datos fueron anonimizados y tratados bajo estrictos criterios de confidencialidad institucional.

C. Preprocesamiento de datos

Antes de la implementación de los modelos, se desarrolló un preprocesamiento de datos, con el objetivo de garantizar la calidad, consistencia y adecuación de la información para el análisis de la retención estudiantil. Este procedimiento se estructuró en las siguientes etapas principales:

Limpieza de datos. Se identificaron y eliminaron registros duplicados. Los valores faltantes fueron tratados mediante estrategias específicas: imputación por la mediana en vari-

ables numéricas y por la moda en variables categóricas, o eliminación del registro cuando la ausencia de datos era insalvable. Adicionalmente, se realizó la detección de valores atípicos mediante análisis exploratorio. Los casos que correspondían a errores de captura fueron corregidos o eliminados, mientras que aquellos que representaban variabilidad natural del fenómeno fueron conservados.

Codificación de variables categóricas. Todas las variables categóricas fueron transformadas a formato numérico. Se empleó label encoding y one-hot encoding según las necesidades del algoritmo a aplicar. Por ejemplo, variables como Modalidad, Facultad y TipoIngreso se convirtieron en variables dummy para evitar introducir relaciones ordinales inexistentes y optimizar su manejo por parte de los modelos. Normalización de variables numéricas. Las variables numéricas, como Edad, CUM, Aprobadas y Reprobadas, fueron estandarizadas mediante el método Z-score (media cero y desviación estándar uno).

División de datos. El conjunto de datos fue segmentado en dos subconjuntos: 75% para entrenamiento y 25% para prueba, manteniendo la proporción original de clases de la variable objetivo mediante stratified sampling. Esta estrategia aseguró una representación equitativa de las categorías en ambos subconjuntos, permitiendo una evaluación objetiva del rendimiento de los modelos.

Este preprocesamiento proporcionó una base sólida para la etapa de modelado, asegurando que los datos estuvieran libres de inconsistencias, en un formato adecuado y listos para maximizar el desempeño de las técnicas empleadas.

D. Análisis descriptivo

El análisis descriptivo constituyó la primera fase del tratamiento estadístico de la información y tuvo como finalidad caracterizar de manera integral el conjunto de datos utilizados en el estudio. Este procedimiento permitió comprender la estructura, distribución y comportamiento de las variables relevantes antes de aplicar los modelos predictivos, proporcionando un contexto sólido para la interpretación de los resultados posteriores.
Las técnicas de análisis descriptivo empleadas incluyeron medidas de tendencia central (media, mediana, moda), medidas de dispersión (desviación estándar, rango intercuartílico) y análisis de frecuencias para variables categóricas. Asimismo, se implementaron visualizaciones gráficas con el propósito de facilitar la interpretación de patrones y tendencias.
Este análisis permitió identificar variables potencialmente influyentes en la retención estudiantil, así como establecer hipótesis iniciales sobre las relaciones entre factores académicos, socioeconómicos y demográficos. Los resultados obtenidos sirvieron como base para el análisis exploratorio y la posterior construcción de los modelos predictivos.

E. Análisis exploratorio: aplicación de metodos no supervisado

Se aplicaron técnicas estadísticas y visualizaciones para describir la distribución de estudiantes en variables clave. Con el objetivo de identificar patrones y reducir la dimensionalidad, se utilizó el Análisis de Componentes Principales (PCA) como técnica preliminar para transformar las variables numéricas y categóricas codificadas en un conjunto reducido de componentes no correlacionados entre sí, que explicaron la mayor parte de la varianza observada. Este procedimiento facilitó la visualización de los datos y optimizó el proceso de agrupamiento posterior al disminuir redundancias y ruido en las variables originales.

En la etapa de segmentación, se emplearon algoritmos de K-medias y agrupamiento jerárquico. El algoritmo de K-medias fue seleccionado como técnica principal debido a su eficiencia y simplicidad para trabajar con grandes volúmenes de datos numéricos estandarizados, permitiendo identificar grupos de estudiantes con patrones similares en variables académicas, demográficas y socioeconómicas, y minimizando la varianza intraclúster.

Como complemento, se utilizó la clusterización jerárquica con un enfoque aglomerativo, generando dendrogramas que facilitaron la exploración de la estructura de los datos y la validación de la coherencia de los grupos obtenidos mediante K-medias. Esta técnica aportó una visión jerárquica de las relaciones entre perfiles estudiantiles sin necesidad de fijar un número previo de clústeres.

La determinación del número óptimo de clústeres se realizó mediante el método del codo, identificando el punto de inflexión. Inicialmente, se estableció k = 2 para una segmentación básica que diferenciara perfiles de mayor y menor riesgo de deserción. Posteriormente, un segundo análisis determinó que k = 5 ofrecía una representación más precisa de la diversidad de perfiles estudiantiles, capturando patrones más específicos y significativos.

F. Análisis predictivo: aplicación de metodos supervisados

Con el propósito de estimar la retención estudiantil, se implementaron y compararon cinco modelos de aprendizaje automático supervisado: Máquinas de Vectores de Soporte (SVM), K-Vecinos Más Cercanos (KNN), Bosques Aleatorios (Random Forest), AdaBoost y XGBoost. La selección de estos algoritmos se basó en su eficacia reportada en la literatura para tareas de clasificación binaria, su capacidad para manejar datos con características mixtas y relaciones no lineales, así como su robustez frente a conjuntos de datos desbalanceados (Chen & Guestrin, 2016; Fernández-Delgado et al., 2014).

Todos los modelos fueron entrenados y optimizados sobre el conjunto de datos de entrenamiento, aplicando búsqueda en malla (Grid Search) y validación cruzada estratificada para ajustar sus hiperparámetros.

- **Máquinas de Vectores de Soporte (SVM):** Se implementó con función kernel para capturar relaciones no lineales, lo que resultó adecuado para entornos de alta dimensionalidad y problemas de clasificación binaria (Peng et al., 2020; Vandamme et al., 2017). Se ajustaron parámetros como el tipo de kernel, el parámetro de regularización (C) y el coeficiente de ajuste (gamma) para maximizar la capacidad de generalización.
- **K-Vecinos Más Cercanos (KNN):** Basado en la similitud entre instancias mediante distancia euclidiana, se evaluaron diferentes valores de k y métodos de búsqueda (ball_tree, kd_tree y brute), seleccionando la configuración óptima mediante validación cruzada (Qingtao et al., 2019).
- **Bosques Aleatorios (Random Forest):** Se entrenó un conjunto de árboles de decisión con muestreo bootstrap y selección aleatoria de características, lo que permitió manejar relaciones complejas, reducir el sobreajuste y estimar la importancia de las variables (Breiman, 2001).
- **AdaBoost (Adaptive Boosting):** Se implementó combinando clasificadores débiles (árboles poco profundos) de forma secuencial, aumentando el peso de las instancias mal clasificadas para mejorar la precisión global, particularmente en entornos con ruido moderado (Schapire & Freund, 2012).
- **XGBoost (Extreme Gradient Boosting):** Se aplicó con técnicas de regularización, poda temprana de árboles y paralelización, configurando un aprendizaje secuencial optimizado para datos estructurados (Chen & Guestrin, 2016).

G. Evaluación del rendimiento de los modelos predictivos

El rendimiento de los modelos se evaluó en el conjunto de prueba mediante métricas específicas para cada modelo predictivo. Estas métricas permitieron obtener una visión integral del desempeño más allá de la simple precisión y facilitaron la comparación entre algoritmos.

Se emplearon las siguientes medidas:

- **Exactitud (Accuracy):** Porcentaje de predicciones correctas sobre el total de observaciones.
- **Precisión (Precision):** Proporción de verdaderos positivos entre todas las predicciones positivas.
- **Sensibilidad o Recall:** Proporción de verdaderos positivos entre todos los casos realmente positivos; en este estudio, representa la capacidad de identificar a estudiantes en riesgo de abandonar (Powers, 2011).
- **Error global:** Complemento de la exactitud, que indica la tasa de errores cometidos por el modelo.
- **Área bajo la curva ROC (AUC-ROC):** Mide la capacidad del modelo para discriminar entre clases en todos los umbrales de decisión, siendo un valor cercano a 1 indicativo de excelente discriminación (Fawcett, 2006).
- **Matriz de confusión:** Permite visualizar verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos, brindando un análisis detallado de los errores de clasificación (Chicco & Jurman, 2020).
-

Este esquema de evaluación garantizó la selección del modelo con mayor capacidad predictiva y menor riesgo de sobreajuste, asegurando la validez de los resultados de predicción de retención estudiantil.

H. Objetivos del estudio

La metodología implementada permitió avanzar de manera estructurada hacia el cumplimiento de los siguientes objetivos:

- Caracterizar la población estudiantil mediante análisis descriptivos, con el fin de identificar la distribución de estudiantes según las características de cada variable demográfica, socioeconómica y académica.

- Identificar patrones y perfiles de retención y deserción estudiantil a través de técnicas de análisis exploratorio, incluyendo métodos de reducción de dimensionalidad y agrupamiento.

- Evaluar y comparar el desempeño de modelos de aprendizaje automático supervisados para la predicción de la retención estudiantil, utilizando métricas de rendimiento que permitieron determinar los modelos más efectivos.

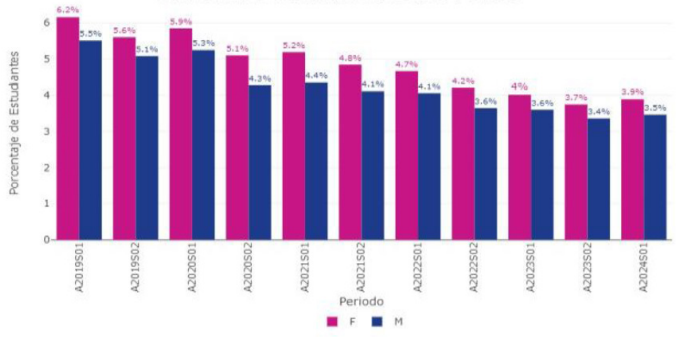
RESULTADOS

El análisis descriptivo constituye la primera etapa para comprender las características esenciales de la población estudiantil antes de la aplicación de técnicas de modelado predictivo. Su propósito es ofrecer una visión integral de las tendencias, patrones y distribuciones de variables clave que inciden en la continuidad académica, así como identificar relaciones preliminares que orienten la selección de atributos relevantes para los modelos.

El estudio de variables demográficas, como edad y género, permite evidenciar diferencias y tendencias asociadas a la retención. En la Fig 1.a, se observa que la mayoría de los estudiantes en cada semestre corresponde al género femenino, acompañado de una tendencia decreciente en el número total de matriculados por periodo. Asimismo, la Fig 1.b muestra que la mayor concentración de estudiantes se encuentra en el rango de 18 a 38 años, manteniéndose esta distribución de manera consistente a lo largo de los semestres.

Figura 1

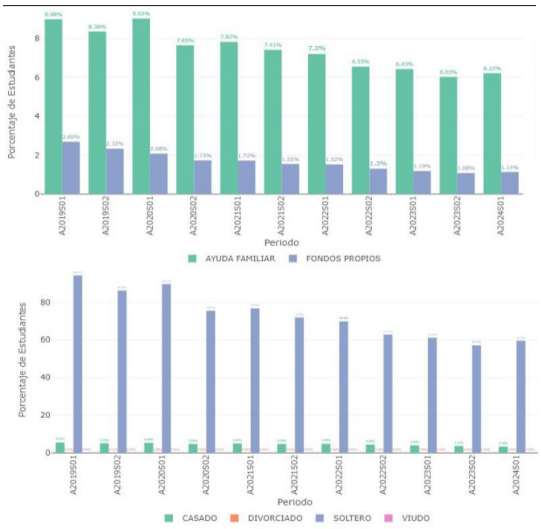
a) Distribución porcentual de estudiantes por género y periodo. b) Distribución de estudiantes por edad y periodo Construcción



En cuanto a las variables socioeconómicas, el análisis revela su influencia directa sobre la permanencia estudiantil. La Fig 2.a indica que la principal fuente de financiamiento proviene del apoyo familiar, mientras que la Fig 2.b, tomando como referencia el total de estudiantes del semestre I-2019, evidencia una disminución progresiva en la matrícula, alcanzando en el semestre I-2024 una reducción significativa respecto al periodo base.

Figura 2

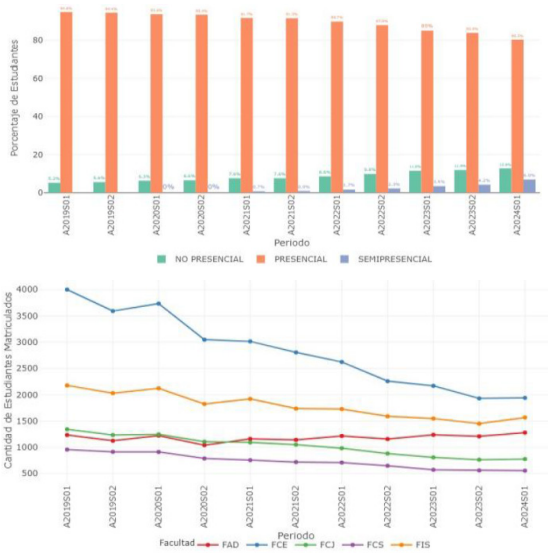
a) Distribución porcentual de estudiantes por tipo de financiamiento y periodo. b) Distribución porcentual de estudiantes por estado familiar y periodo



El análisis de las variables académicas profundiza en cómo factores institucionales, como la modalidad de estudio y la estructura de oferta académica, condicionan la trayectoria de los estudiantes. La Fig 3.a presenta la distribución de estudiantes según modalidad (presencial, semipresencial y no presencial) por periodo, y la Fig 3.b muestra la tendencia de matrícula por facultad a lo largo de los ciclos académicos, evidenciando variaciones diferenciales entre áreas de formación.

Figura 3

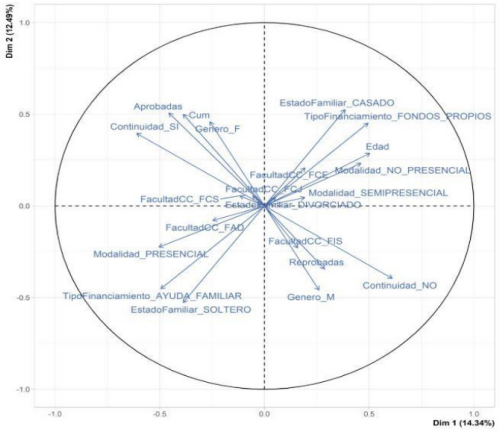
a) Distribución de estudiantes por modalidad y periodo. b) Tendencia de matrícula por periodo y facultad



El análisis exploratorio se complementó con la aplicación del Análisis de Componentes Principales (PCA), utilizado como etapa preliminar a la clusterización. Para ello, las variables cualitativas fueron transformadas en dicotomicas, lo que permitió representar de forma simultánea la variabilidad de todas las variables en un espacio de menor dimensionalidad. El biplot resultante (Fig 4) revela la estructura de los datos y facilita la identificación visual de patrones de agrupamiento iniciales.

Figura 4

a) Biplot de Análisis de Componentes Principales (PCA)



Como primer ejercicio, se ejecutó un análisis de clústeres con k=2, obteniéndose la distribución porcentual de estudiantes por facultad para cada grupo (Fig 5). No obstante, con el fin de determinar un valor óptimo de k clusters que mejore la representatividad y la complejidad del modelo, se aplicó el método del codo (Fig 6). Este procedimiento mostró una inflexión significativa en k=5, lo que sugiere que este número de grupos captura adecuadamente la heterogeneidad de la población sin incurrir en sobreajuste.

Figura 5

a) Porcentaje de individuos de cada clúster por Facultad para k=2

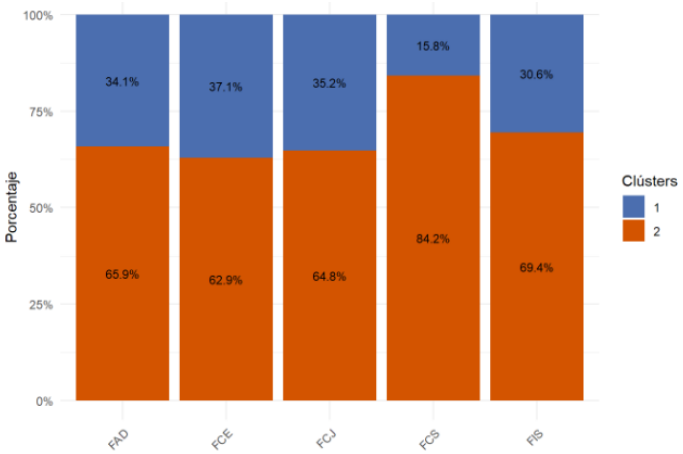
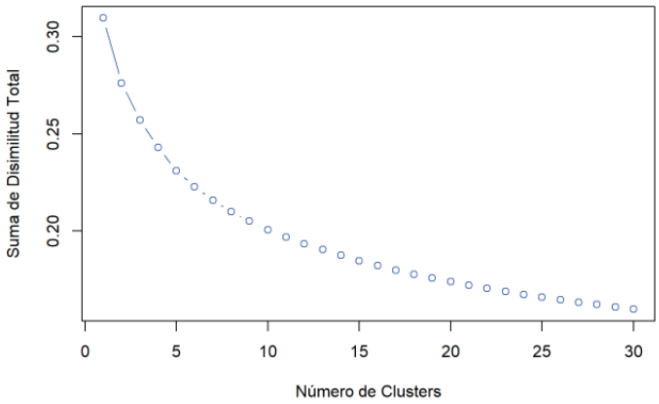


Figura 6

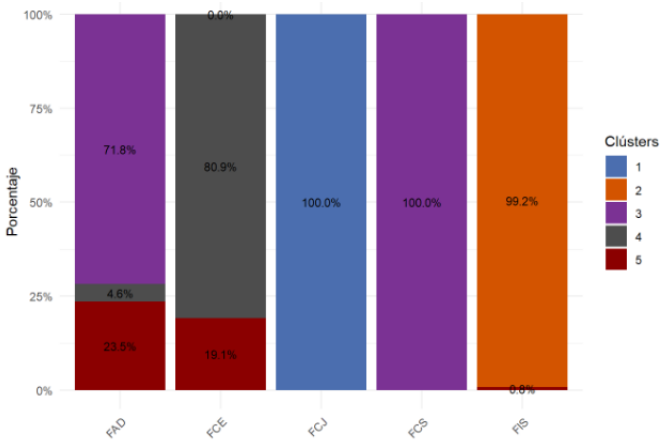
Aplicación del método del codo para determinar el número óptimo de clústers



La segmentación final con k=5 (Fig 7) permitió identificar perfiles diferenciados de estudiantes según la facultad, revelando patrones específicos asociados a factores académicos. Estos hallazgos exploratorios constituyeron la base para la siguiente etapa del análisis, orientada a la construcción y evaluación de modelos predictivos con el propósito de optimizar la estimación de la retención estudiantil.

Figura 7

Porcentaje de individuos de cada clúster por Facultad para k=5



Finalmente, se entrenaron y evaluaron diversos algoritmos supervisados, seleccionando aquellos con el mejor desempeño para su análisis comparativo. Entre los modelos con resultados más destacados se encuentran el Máquinas de Vectores de Soporte (SVM) optimizado, el método de K vecinos más cercanos (implementado con algoritmos de búsqueda ball_tree, kd_tree y brute), así como Bosques Aleatorios, XGBoost y AdaBoost. La Tabla I muestra el

rendimiento de estos modelos, permitiendo contrastar su capacidad predictiva y su aplicabilidad en entornos institucionales de educación superior.

Tabla 1

Desempeño de Modelos Predictivos de Retención Estudiantiltimo de clústers

Modelo	Parámetros optimizados	Precisión global
SVM	kernel=rbf C=46	77.31%
KNN	algorithm=ball_tree n_neighbors=9	75.74%
Random Forest	entropy n_estimators=50 min_samples_split=32	78.99%
XGBoost	n_estimators=100 min_samples_split=31	78.91%
AdaBoost	entropy n_estimators=50 min_samples_split=29	76.21%

DISCUSIÓN

El análisis demográfico evidencia que la población estudiantil está compuesta mayoritariamente por mujeres en todos los periodos analizados, pese a la tendencia decreciente en la matrícula total. Este grupo presenta una probabilidad de continuidad superior a la de los hombres, lo que sugiere una mayor permanencia académica. Asimismo, los rangos etarios de 18-24 y 25-34 años concentran la mayor proporción de estudiantes y constituyen segmentos clave para la implementación de estrategias de retención. Estos hallazgos refuerzan la relevancia de los factores demográficos como determinantes en la permanencia estudiantil y orientan la necesidad de intervenciones focalizadas.

Desde la perspectiva socioeconómica, se observa una disminución general en la matrícula, con predominio de estudiantes que financian sus estudios mediante apoyo familiar. Aquellos que dependen exclusivamente de recursos propios, aunque representan un grupo menor, exhiben mayores probabilidades de abandono entre ciclos, pero también una tasa de graduación más elevada. La reducción de matrícula afecta tanto a estudiantes solteros como casados, siendo los primeros los más impactados. La variabilidad en la continuidad según el estado familiar evidencia que las condiciones económicas y contextos personales influyen significativamente en la retención y finalización de los estudios, lo que destaca la necesidad de políticas institucionales que contemplen estos factores.

En el ámbito académico, la modalidad presencial sigue siendo la predominante, aunque ha registrado una disminución notable en la matrícula desde 2019, mientras que la modalidad no presencial ha crecido de forma sostenida. Este cambio refleja una adaptación hacia esquemas educativos más flexibles, probablemente influenciada por el avance de la digitalización y factores externos recientes. Sin embargo, la modalidad presencial presenta mayores tasas de deserción, lo que demanda intervenciones específicas para este grupo. A nivel de facultades, la Facultad FCE enfrenta los mayores desafíos de retención, mientras que la Facultad FAD y la Facultad FCJ mantienen niveles estables de continuidad, lo que respalda la necesidad de estrategias diferenciadas por área académica.

El análisis por nivel académico confirma que los estudiantes de primeros niveles presentan mayores índices de abandono, mientras que en niveles avanzados la continuidad es más frecuente, lo que sugiere un incremento del compromiso académico conforme progresa la trayectoria formativa. Estos hallazgos proporcionan un panorama integral del progreso académico y ofrecen insumos clave para el diseño de estrategias de permanencia.

Los resultados en el análisis descriptivo justifican la aplicación de un análisis exploratorio focalizado en los estudiantes de primer nivel. Este procedimiento, basado en la segmentación no supervisada de la población en grupos homogéneos, permite identificar patrones latentes sin necesidad de etiquetas previas, lo que posibilita el desarrollo de estrategias de intervención personalizadas y orientadas a las necesidades específicas de cada segmento estudiantil.

La primera etapa de clusterización, con k = 2, proporcionó una división general de la población estudiantil en dos grupos bien diferenciados. Aunque esta configuración permitió detectar diferencias notables, su baja granularidad motivó la exploración de valores superiores de k. La aplicación del método del codo determinó que k = 5 ofrecía el mejor equilibrio entre variabilidad explicada e interpretabilidad, evitando el sobreajuste y capturando con mayor precisión la heterogeneidad de los datos.

La segmentación con k = 5 reveló combinaciones específicas de factores demográficos, socioeconómicos y académicos, generando información valiosa para el diseño de intervenciones focalizadas. El uso de Análisis de Componentes Principales (PCA) resultó clave para reducir la dimensionalidad de los datos, facilitar la visualización bidimensional de los clústeres y destacar las variables más influyentes en su formación, optimizando así el rendimiento del proceso de agrupamiento.

En la fase predictiva, se evaluaron diversos algoritmos de

aprendizaje automático incluyendo SVM con kernel RBF, KNN (con ball_tree, kd_tree y brute), Bosques Aleatorios, XGBoost y AdaBoost para determinar el mejor desempeño en términos de precisión global, positiva y negativa. El SVM con kernel RBF y C = 46 alcanzó un rendimiento notable (77.31 % de precisión global) con un adecuado balance en la clasificación de ambas clases, aspecto crítico en la predicción de retención.

No obstante, el mejor resultado fue obtenido por el modelo de Bosques Aleatorios (entropy, n_estimators

= 50, min_samples_split = 32), que alcanzó una precisión global del 78.99 %. Este algoritmo destacó por su alta precisión negativa, evidenciando una capacidad superior para identificar a los estudiantes con mayor probabilidad de continuar sus estudios. Si bien los modelos KNN y los métodos de ensamble como XGBoost y AdaBoost también mostraron un desempeño aceptable, se estima que podrían alcanzar mejores resultados mediante una calibración más exhaustiva de hiperparámetros (ver Tabla I).

En resumen, estos hallazgos demuestran que la combinación de análisis exploratorio no supervisado y modelado predictivo supervisado constituye una estrategia sólida para la identificación temprana de perfiles de riesgo y la optimización de intervenciones en el ámbito de la retención estudiantil.

CONCLUSIONES

La investigación confirma que la combinación de técnicas de análisis exploratorio no supervisado y modelos predictivos supervisados de aprendizaje automático ofrece una visión integral y operativa de los factores que inciden en la retención estudiantil. La aplicación de clusterización permitió segmentar a la población en perfiles homogéneos de alto y bajo riesgo de deserción, facilitando el diseño de estrategias de intervención personalizadas. También, los modelos supervisados alcanzaron un alto nivel de desempeño, demostrando su potencial para ser incorporados en sistemas de gestión académica con el fin de generar alertas tempranas y habilitar acciones preventivas oportunas.

El análisis reveló que variables como el rendimiento académico, la modalidad de estudio y el tipo de financiamiento tienen un peso significativo en la continuidad académica. En particular, los estudiantes con bajo CUM, modalidad presencial y financiamiento propio conforman el grupo más vulnerable a la deserción, por lo que se recomienda priorizar intervenciones específicas para este segmento.

Finalmente, se propone ampliar el alcance de la investigación mediante la incorporación de variables de carácter psicopedagógico y el empleo de técnicas avanzadas como el aprendizaje profundo (Deep Learning), con el objetivo de incrementar la capacidad explicativa y mejorar la precisión en la predicción del abandono estudiantil.

REFERENCIAS

Abdullah, N., Mohamad, A., & Bakar, A. (2019). Predicting student dropout using clustering techniques.

International Journal of Advanced Computer Science and Applications, 10(9), 73-81.

Arqawi, S., Zitawi, E., Rabaya, A., Abunasser, B., & Abu-Naser, S. (2022). Predicting University Student Retention using Artificial Intelligence. International Journal of Advanced Computer Science and Applications. https://doi.org/10.14569/ijacsa.2022.0130937.

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. Computers & Education, 113, 177-194.

Astin, A. (1999). Student involvement: A developmental theory for higher education. Journal of College Student Development, 40(5), 518-529.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

Cabrera, A., Nora, A., & Castaneda, M. (2006). College persistence: Structural equations modeling test of an integrated model of student retention. The Journal of Higher Education, 64(2), 123-139.

Cabrera, A., Nora, A., & Castaneda, M. (2018). The role of finances in the persistence process: A structural model. Research in Higher Education, 33(5), 571-593.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 785-794. https://doi.org/10.1145/2939672.2939785

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21(6), 1-13. https://doi.org/10.1186/s12864-019-6413-7

Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. In Proceedings of the 5th Annual International Conference on Educational Data Mining, 5-12.

Fahd, K., Venkatraman, S., Miah, S. J., & Ahmed, K. (2022). Application of machine learning in higher education to assess student academic performance, at-risk, and attrition: A meta-analysis of literature. Education and Information Technologies, 1-33.

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? Journal of Machine Learning Research, 15(1), 3133- 3181. Obtenido de http://jmlr.org/papers/v15/delgado14a.html

Gupta, S., Lehmann, D., & Stuart, J. (2006). Valuing customers. Journal of Marketing Research, 43(1), 7-18.

Hagedorn, L. S. (2021). How to define retention: A new look at an old problem. Journal of College Student Retention: Research, Theory & Practice, 23(1), 1-15.

Herzog, S. (2006). Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression. New Directions for Institutional Research(131), 17-33.

Kotler, P., & Armstrong, G. (2018). Principles of Marketing (17th ed.). Education Pearson. Kotler, P., & Keller, K. (2016). Marketing Management (15th ed.). Pearson Education.

Kurniawan, T., & Taufiq, M. (2019). Clustering students based on their academic data using K-Means algorithm. Journal of Information Systems Engineering and Business Intelligence, 5(2), 110-118. Obtenido de https://www.ijstr.org/final-print/aug2019/Clustering-Student-Data-Based-On-K-means-Algorithms.pdf

Martínez, R., Caballero, A., & Figueroa, R. (2021). Early detection of student dropout in higher education using machine learning techniques. Education and Information Technologies, 26, 2133-2152. Obtenido de https://www.researchgate.net/publication/372493356_Predicting_Student_Dropout_

Meneses, J., Rodríguez-Gómez, D., & Fernández, C. (2020). Determinants of higher education drop-out in Latin America: A systematic review. Journal of Latin American Studies, 52(3), 645-665.

Peng, L., Chaoli, S., Guochen, Z., & Yaochu, J. (2020). Multi-surrogate multi-tasking optimization of expensive problems. Knowledge-Based Systems. https://doi.org/10.1016/j.knosys.2020.106262

Powers, D. M. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, 2(1), 37-63.

Qingtao, X., Xin, Z., & Chenghua, Z. (2019). Application Research of KNN Algorithm Based on Clustering in Big Data Talent Demand Information Classification. International Journal of Pattern Recognition and Artificial Intelligence, 34(19). https://doi.org/10.1142/S0218001420500159

Romero, C., Ventura, S., & García, E. (2013). Data mining in course management systems: Moodle case study and tutorial. Computers & Education, 51(1), 368-384.

Schapire, R., & Freund, Y. (2012). Boosting: Foundations and Algorithms. MIT Press (MA). https://doi.org/10.7551/mitpress/8291.001.0001

Simões, A., & Soares, M. (2019). A model of student attrition and retention in higher. Educational Research and Evaluation, 25(2-3), 25(2-3), 133-152.

Thomas, L. (2018). Understanding student retention and engagement: Contexts and strategies for improvement. Higher Education Quarterly, 72(3), 278-289.

Tianqi, C., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Association for Computing Machinery(10), 785–794. https://doi.org/10.1145/2939672.2939785

Tinto, V. (2012). Completing college: Rethinking institutional action. University of Chicago Press. Tinto, V. (2017). Through the eyes of students. Journal of College Student

Retention: Research, Theory & Practice, 19(3), 254-269.

Vandamme, J.-P., Meskens, N., & Superby, J.-F. (2017). Predicting academic performance by data mining methods. Education Economics, 15(4), 405-419. Obtenido de https://ideas.repec.org/a/taf/edecon/v15y2007i4p405-419.html