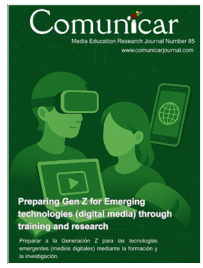
Comunicar, n. 85, v. XXXIV, 2026 | Media Education Research Journal | ISSN: 1134-3478; e-ISSN: 1988-3478
www.comunicarjournal.com



Aplicación de modelos de Machine Learning para la predicción de la deserción escolar en estudiantes de una institución colombiana de formación por competencias

Application of Machine Learning Models for Predicting School Dropout in Students from a Colombian Competency-based Education Institution

John Jairo Castro-Maldonado*, Servicio Nacional de Aprendizaje SENA, Medellín (Colombia) (jcastrom@sena.edu.co) (https://orcid.org/0000-0002-3823-4297)
Jennifer Andrea Londoño-Gallego, Esp. en Formulación y evaluación de proyectos, Servicio Nacional de Aprendizaje SENA, Medellín (Colombia) (jealondonog@sena.edu.co) (https://orcid.org/0000-0003-2957-9178)
Paula Andrea Rodríguez-Marín, Instituto Tecnológico Metropolitano ITM, Medellín (Colombia) (paularodriguez7913@correo.itm.edu.co) (https://orcid.org/0000-0002-3547-560X)
Juan David Martínez-Vargas, Universidad EAFIT, Medellín (Colombia) (jdmartinev@eafit.edu.co) (https://orcid.org/0000-0001-7037-6925)

* Indicates the corresponding author

RESUMEN

La deserción estudiantil constituye un desafío estructural en la educación superior colombiana, especialmente en contextos con sistemas curriculares y pedagógicos rígidos donde resulta complejo implementar estrategias preventivas oportunas. Este estudio desarrolla y valida un modelo híbrido de aprendizaje automático, fundamentado en la metodología CRISP-DM, que combina algoritmos supervisados (*Random Forest*, *Ridge*, *XGBoost*, KNN) y no supervisados (*K-Means*, DECLA), apoyados en técnicas de reducción y segmentación (PCA, ACM). A partir de variables sociodemográficas, indicadores de desempeño académico y un instrumento de seguimiento diseñado ad hoc, los modelos alcanzaron una alta precisión para anticipar el riesgo de abandono y segmentar a los estudiantes en perfiles de alta, media y baja probabilidad de deserción. Los algoritmos basados en árboles, en particular *Random Forest*, evidenciaron el mejor desempeño, identificando predictores críticos como cantidad de quejas, reversiones de calificaciones, estrato socioeconómico, género y estado civil. La principal contribución de este trabajo radica en trasladar la analítica predictiva de un ejercicio experimental hacia un sistema de apoyo institucional en programas de educación superior por competencias, donde la rigidez académica suele limitar la intervención temprana. Al anticipar la deserción mediante evidencia empírica en tiempo real, el modelo permite diseñar rutas diferenciadas de acción: tutorías personalizadas, apoyos socioeconómicos y flexibilización curricular que complementan las reformas educativas de largo plazo. De esta manera, se justifica su relevancia en la educación superior como recurso innovador y fundamentado para fortalecer la permanencia estudiantil.

ABSTRACT

Student dropout is a structural challenge in Colombian higher education, particularly in contexts with rigid curricular and pedagogical systems where the implementation of timely preventive strategies is complex. This study develops and validates a hybrid machine learning model, based on the CRISP-DM methodology, that integrates supervised algorithms (Random Forest, Ridge, XGBoost, KNN) and unsupervised approaches (K-Means, DECLA), supported by dimensionality reduction and segmentation techniques (PCA, MCA). Using sociodemographic variables, academic performance indicators, and a specifically designed monitoring instrument, the models achieved high accuracy in anticipating dropout risk and segmenting students into profiles of high, medium, and low probability of withdrawal. Tree-based algorithms, particularly Random Forest, demonstrated the best performance, identifying critical predictors such as number of

Received: 2025-08-25 | Reviewed: 2025-09-29 | Accepted: 2025-10-05 | Online First: 2026-04-11 | Published: 2026-04-15
DOI: https://doi.org/10.5281/zenodo.19690782 | Pages: 171-202

complaints, grade reversals, socioeconomic status, gender, and marital status. The main contribution of this work lies in moving predictive analytics from an experimental exercise to an institutional support system in competency-based higher education, where academic rigidity often limits early interventions. By anticipating dropout through real-time empirical evidence, the model enables the design of differentiated action pathways personalized tutoring, socioeconomic support, and curricular flexibility that complement long-term educational reforms. In this way, its relevance in higher education is justified as an innovative and evidence-based resource to strengthen student retention.

PALABRAS CLAVE / KEYWORDS
Intención de abandono, Inteligencia Artificial, aprendizaje automático, educación.
Intention to Drop Out, Artificial Intelligence, Machine Learning, Education.

1. Introducción

La deserción es un concepto que ha tenido interpretaciones por parte de diversos autores, inicialmente, se puede indicar que la deserción se puede considerar como un proceso de abandono forzoso o voluntario de un proceso formativo por influencia de elementos internos o externos del estudiante, por otra, parte se menciona que es el abandono prematuro de un programa antes del alcanzar su certificación y considera un tiempo extenso para que el alumno se pueda incorporar (Sanhueza Gutiérrez et al., 2021). En ese sentido, el fenómeno de la deserción se presenta desde hace años afectando diferentes dimensiones de los países y sus sistemas educativos en específico, toda vez, que la reducción de la escolaridad de los ciudadanos de un estado es directamente proporcional a la apropiación y desarrollo de la tecnología en un sociedad globalizada e hiperconectada (Fuertes Arroyo y Uc Ríos, 2023), donde las principales empresas de la actualidad se enfocan en procesos de gestión del conocimiento más que en procesos manufactureros, es así, que hoy en día hablamos de la "mentefactura" más que la manufactura.

Marrón Ramos et al. (2022), hace un análisis sobre diferentes puntos de vista de autores que han estudiado este tema, en el texto se resalta, que la deserción es un tópico que, de no ser tratado por las instituciones de manera efectiva, afectaría no solo las economías de los claustros estudiantiles, sino que afectaría los programas académicos ofertados por las instituciones debido a que limitaría su oferta en años siguientes. Por su parte, la UNICEF estima que aproximadamente una tercera parte de la población en edad escolar no tuvo acceso a la educación durante los años 2020 y 2021 (Caceres-Correa, 2021), en América Latina la tasa de deserción se plantea en el 55% y en Colombia, el porcentaje de estudiantes que se matriculan por primera vez en programas de licenciatura y desertan al finalizar el primer año alcanza el 22%, cifra que supera significativamente el promedio de los países de la OCDE, establecido en 13% (OECD, 2025).

El fenómeno del abandono escolar responde a una combinación de factores endógenos y exógenos. Entre los primeros se destacan variables individuales como la resiliencia, la motivación intrínseca y la inteligencia emocional de los estudiantes. Los factores exógenos, en cambio, están vinculados a elementos externos que influyen de manera sustancial en la decisión de abandonar, tales como la calidad de la práctica pedagógica, la disponibilidad y adecuación de la infraestructura institucional, así como el acompañamiento socioafectivo y de bienestar ofrecido por la institución. En este sentido, las universidades y centros de educación superior tienen la responsabilidad de diseñar e implementar estrategias que fortalezcan la motivación estudiantil, destacando el valor transformador de la educación y su impacto positivo en la calidad de vida de los individuos (Améstica-Rivas et al., 2021; Arnaud Bobadilla et al., 2022; Di Paola Naranjo et al., 2022).

En este marco, la Tabla 1 sistematiza una serie de factores identificados por diversos autores como detonantes del abandono en la educación superior. Aunque las investigaciones revisadas emplean terminologías distintas para describir las conductas o circunstancias asociadas a la deserción, el trasfondo del fenómeno mantiene una notable coherencia. Asimismo, la tabla especifica las fuentes de información utilizadas en cada estudio, diferenciando entre aquellas que recurrieron a metodologías de campo como encuestas, cuestionarios o entrevistas semiestructuradas y las que se basaron en análisis documentales o registros administrativos.

Un aspecto relevante es que solo un número reducido de investigaciones trasciende la identificación de causas para proponer estrategias de intervención orientadas a la prevención del abandono. Esta ausencia de mecanismos de respuesta evidencia una brecha en la literatura, que suele centrarse en el diagnóstico de factores, pero no en el diseño de acciones correctivas o preventivas. La información recopilada en la Tabla 1 permite, por tanto, no solo comprender la multiplicidad de variables que inciden en la deserción, sino también reconocer la necesidad urgente de articular los hallazgos con políticas y programas institucionales que fortalezcan la permanencia estudiantil.

En relación con este tema, la Organización para la Cooperación y el Desarrollo Económicos (OCDE) ha indicado que, a pesar de las mayores oportunidades de acceso a la educación superior para los estudiantes, no se dan las condiciones adecuadas que promuevan el progreso en el plan de estudios, la retención, la calidad de la formación, la graduación oportuna y la integración de los estudiantes más desfavorecidos. Como resultado, las desigualdades y las inequidades continúan siendo perpetuadas en la región, en otras palabras, si bien, puede haber un proceso de matrícula eficaz de los estudiantes hay elementos institucionales que aún perpetúan la problemática de la deserción (Améstica-Rivas et al., 2021).

Actualmente, muchas instituciones dependen en gran medida de métodos tradicionales y reactivos para identificar a los estudiantes en riesgo de deserción. Estos enfoques suelen basarse en señales evidentes de bajo rendimiento académico o la falta de asistencia regular, lo que a menudo significa que la intervención se produce cuando el problema ya está bastante avanzado.

Tabla 1: Motivos de abandono y fuentes de consulta.

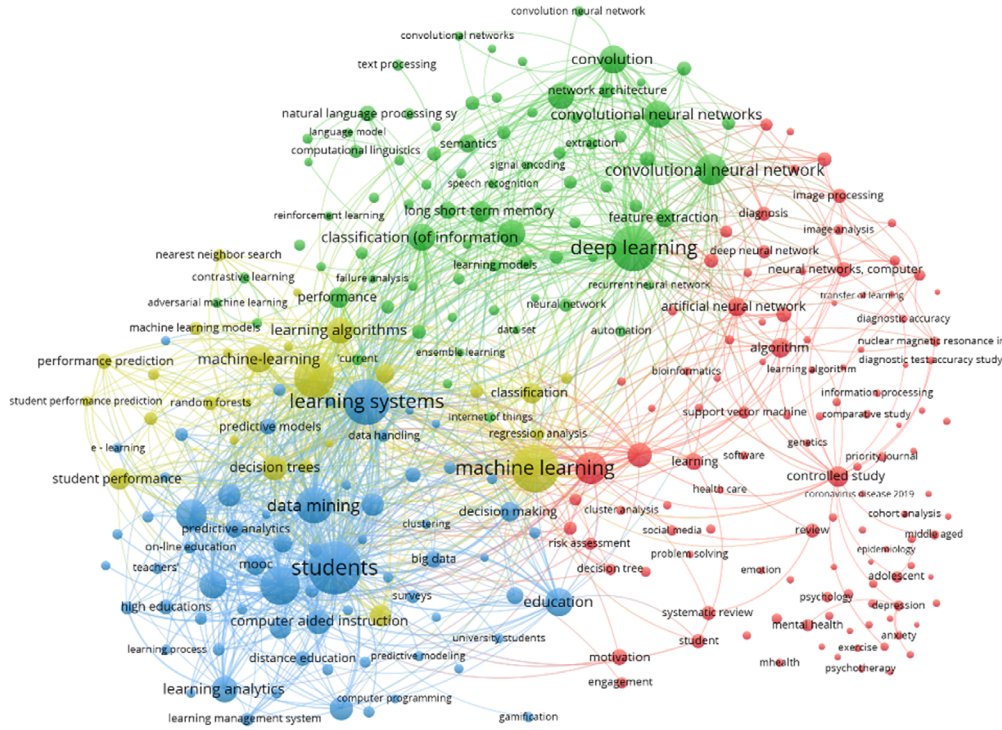
Referencia	Causas o motivos del abandono escolar	Fuentes de información	Existen pruebas de estrategias de prevención del abandono escolar.
(Zimányi et al., 2022)	Falta de apoyo en la titulación. Dificultades en la realización del trabajo de fin de carrera. Desconocimiento de los requisitos de titulación. Problemas económicos. Desmotivación académica. Falta de orientación académica. Necesidad de trabajo a tiempo completo. Dificultades en el aprendizaje de idiomas. Desconocimiento de los programas educativos.	Encuestas, entrevistas semiestructuradas	NO
(Caceres-Correa, 2021)	Falta de acceso a la educación en línea Consecuencias económicas	Datos de UNICEF Datos del Ministerio de Educación de Chile, la SEP y el Ministerio de Educación de México.	NO
(de Freitas e Silva y Bezerra Sampaio, 2022)	Falta de apoyo financiero Bajo rendimiento académico Autopercepción Políticas institucionales	Análisis de documentos (documentos políticos)	NO
(Sanhueza Gutiérrez et al., 2021)	La infraestructura (metros cuadrados de salas, laboratorios, bibliotecas) afecta al desgaste. El número de bibliotecas académicas influye en la tasa de abandono. Número de profesores con postgrado	Análisis de datos de 28 universidades chilenas (12 privadas, 16 estatales) para comparación.	NO
(Pereira Santana y Vidal Cortez, 2020)	Falta de vinculación con la institución. Problemas de motivación, como falta de interés y desesperanza. Factores económicos y personales, como dificultades financieras. Cambio de vocación o carrera. Problemas académicos, como bajo rendimiento. Falta de orientación y apoyo institucional. Desafíos psicológicos y emocionales. Cambios en la vida personal.	Análisis de la documentación científica	NO
(Navarro Roldán y Zamudio Sisa, 2021)	Estructura de apoyo al plan de estudios e inserción laboral Autoeficacia y decisión profesional Redes de apoyo funcionales y disfuncionales	Cuestionario de abandono universitario para estudiantes (CDUe) basado en la Teoría Ecológica del Desarrollo Humano, que incluye seis escalas con 63 ítems y una escala de respuesta tipo Likert. También se utilizó el Cuestionario de Experiencias Académicas, versión reducida (QVA-r) con 60 ítems y opciones de respuesta tipo Likert.	NO
(Fuertes Arroyo y Uc Ríos, 2023)	Falta de motivación debido a una metodología de enseñanza poco estimulante. Escasez de profesionales en el campo de las tecnologías de la información en el país. La necesidad de que los profesores se adapten a las exigencias de la era tecnológica. Mayor interés de los estudiantes por la educación mediada por la tecnología.	Encuesta validada con prueba de validez y confiabilidad Alfa de Cronbach aplicada a 81 estudiantes de carreras afines a las tecnologías en Medellín, Colombia, con muestreo probabilístico.	NO
(Lozano Treviño y Maldonado Maldonado, 2022)	Reprobación Aspectos positivos Profesores Orden institucional Características positivas de la institución Recursos institucionales Tasa de abandono	Cuestionario estandarizado para estudiantes.	NO

Tabla 1: Motivos de abandono y fuentes de consulta.

Referencia	Causas o motivos del abandono escolar	Fuentes de información	Existen pruebas de estrategias de prevención del abandono escolar.
(Chalpartar Nasner et al., 2022)	Factores familiares, especialmente la situación económica, que limita el acceso a los dispositivos electrónicos y a las redes de Internet. Necesidad de acceso a Internet y dispositivos electrónicos para la formación académica virtual. Importancia del apoyo económico y moral de la familia en la permanencia del alumno. Necesidad de estructurar estrategias de orientación, apoyo y acompañamiento de los estudiantes y sus familias. Nuevas demandas de competencias informáticas para profesores y alumnos en el entorno digital.	Encuesta y entrevista semiestructurada.	SÍ
(Marrón Ramos et al., 2022)	Gestión universitaria Vocación y apoyo Demanda y compañía	Entrevista semiestructurada con los estudiantes	NO
(Hernández-Medina y Ramírez-Torres, 2022)	Impacto del programa de ayuda financiera. Variables de rendimiento académico.	Diseño de discontinuidades de regresión difusa con cuatro puntos de corte en el índice de ayuda. Encuestas y entrevistas semiestructuradas.	SÍ
(Di Paola Naranjo et al., 2022)	Madre con experiencia en la enseñanza superior. Vivir solo Haber cursado estudios secundarios de forma continuada. Edad entre 19 y 25 años (por desgaste)	Cuestionario sociodemográfico autoadministrado. Expedientes administrativos de rendimiento académico	NO
(Cabrales et al., 2022)	Causas multifactoriales y complejas: personales, sociales y sanitarias.	Registros de la oficina de admisiones Bases de datos de la oficina de planificación Sistema de Prevención y Análisis de la Deserción en Instituciones de Educación Superior (SPADIES)	NO
(Arnaud Bobadilla et al., 2022)	Malos hábitos de estudio. Falta de preparación para el bachillerato. Escasa capacidad de redacción y comprensión lectora. Uso predominante de clases magistrales. Deficiencias en razonamiento lógico, orientación profesional, expectativas, motivación, capacidad de comunicación oral y escrita, hábitos de estudio y disciplina. Expectativas individuales, motivación y autoestima.	Análisis de documentos. Encuestas y grupos de discusión.	SÍ
(Améstica-Rivas et al., 2021)	Pérdida de la carrera profesional debido al rendimiento académico Dimisión definitiva por motivos personales Matrícula y tasas	Datos académicos facilitados por la institución (tasas académicas, derechos de matrícula, etc.)	NO

Ante este panorama de inequidades persistentes y respuestas institucionales tardías, surge la necesidad de explorar enfoques innovadores que trasciendan las limitaciones de los métodos tradicionales. En este sentido, la literatura reciente muestra un creciente interés en el uso de la inteligencia artificial y, particularmente, de los modelos de *Machine Learning* como herramientas para anticipar riesgos de deserción antes de que se materialicen. Con el fin de analizar esta tendencia, se realizó una búsqueda sistemática con la ecuación *"Education" AND ("Artificial Intelligence" OR "Machine Learning") AND ("Desertion" OR "Dropout")* y se aplicó un análisis bibliométrico con VOSviewer (versión 1.6.16). Los resultados, representados en la Figura 1, evidencian que *"Machine Learning"* se constituye en el nodo central de las investigaciones recientes, en estrecha relación con conceptos como *"Students"*, *"Learning Systems"* y *"Deep Learning"*. Esto confirma que la comunidad científica está priorizando enfoques basados en datos y algoritmos como una alternativa viable para superar la dependencia de intervenciones reactivas, ofreciendo nuevas posibilidades de predicción temprana y de diseño de estrategias de retención más efectivas (Hoca y Dimililer, 2025; Martínez y Castillo, 2024).

Figura 1: Red nodal de conceptos encontrados en la literatura con el mayor número de co-ocurrencias.



Partiendo de esta base, la Tabla 2 presenta una selección de estudios sobre la aplicación de herramientas específicas de inteligencia artificial (IA) en procesos educativos. Esta síntesis evidencia las técnicas más empleadas como el procesamiento del lenguaje natural (PLN), la visión por computador, el aprendizaje profundo y los sistemas de tutoría inteligentes, sus aplicaciones en diferentes ámbitos de la enseñanza, y si a partir de ellas se han propuesto o no estrategias concretas de prevención del abandono escolar.

El análisis comparativo revela un hallazgo significativo: aunque la literatura demuestra un crecimiento sostenido en la incorporación de IA para tareas como la tutoría virtual, la retroalimentación personalizada, la generación automática de contenidos y el apoyo en la enseñanza de disciplinas específicas, no se evidencian propuestas sólidas de estrategias institucionales orientadas a reducir la deserción escolar. En la mayoría de los casos, las aplicaciones de IA se han centrado en mejorar aspectos didácticos, instruccionales o recreativos, sin traducirse en mecanismos integrales de retención académica (Bustamante Bula y Camacho Bonilla, 2024; Forero-Corba y Negre Bennasar, 2024).

Este vacío en la investigación subraya la necesidad de ampliar el alcance de los modelos de *Machine Learning* y otras técnicas de IA hacia dimensiones administrativas y de gestión educativa. Incorporar estos modelos a los sistemas institucionales de alerta temprana permitiría no solo predecir el riesgo de abandono, sino también diseñar intervenciones específicas y adaptadas al contexto pedagógico de cada institución. En consecuencia, la integración de la analítica predictiva con políticas de acompañamiento académico, programas de bienestar y apoyos socioeconómicos podría potenciar de manera significativa el impacto de estas herramientas en la permanencia estudiantil y en los resultados académicos generales (Bustamante Bula y Camacho Bonilla, 2024; Rodríguez Almazán et al., 2023).

La capacidad de identificar y predecir patrones de deserción estudiantil representa una diferencia sustancial frente a los enfoques tradicionales. Mientras que las reformas pedagógicas y curriculares requieren largos procesos de diseño, implementación y validación institucional, el uso de modelos de aprendizaje automático permite anticipar riesgos de abandono con base en información empírica en tiempo real. Esta anticipación facilita la toma de decisiones preventivas, orientadas a fortalecer la retención antes de que el

estudiante abandone el sistema. En este sentido, recientes investigaciones han mostrado que la integración de algoritmos predictivos en entornos educativos incrementa la eficacia de los programas de permanencia en un 20% al permitir intervenciones tempranas (Hoca y Dimililer, 2025; Martínez y Castillo, 2024).

Tabla 2: Herramientas de Inteligencia Artificial aplicadas a la Educación.

Referencia	Técnicas utilizadas	Aplicación en la educación	Existen pruebas de estrategias de prevención del abandono escolar.
(Rodríguez Chávez, 2021)	Lógica difusa Razonamiento basado en casos Agentes inteligentes Red neuronal artificial Redes bayesianas Lingüística difusa Representación del conocimiento Lingüística computacional Procesamiento del lenguaje natural (PLN) Visión artificial	Sistemas de tutoría inteligentes (STI)	NO
(Ocaña-Fernández et al., 2019)	Reconocimiento automático del habla (ASR) Procesamiento del lenguaje natural (PLN)	Sistemas de tutoría inteligentes (STI)	NO
(Llanos Mosquera et al., 2021)	Procesamiento del lenguaje natural (PLN) Generación de lenguaje natural Visión artificial	Ciberaulas evaluación automática del código	NO
(Vidal Ledo et al., 2019)	Procesamiento del lenguaje natural (PLN)	Tutores inteligentes Sistemas de gestión del aprendizaje	NO
(Coto Jiménez, 2021)	Aprendizaje profundo Automatización robótica Procesamiento del lenguaje natural (PLN) Aprendizaje automático Minería de datos Redes neuronales Computación evolutiva Sistemas basados en reglas Reconocimiento de patrones Sistemas expertos Lógica difusa	Laboratorios inteligentes	NO
(Hidalgo Suarez et al., 2023)	Creación de API Procesamiento del lenguaje natural (PLN) Clasificación de algoritmos Aprendizaje profundo	Jueces virtuales Aprendizaje analítico Chatbots para el aprendizaje Universidad 4.0 Plataformas inteligentes Sistemas educativos adaptativos Sistemas de tutoría inteligentes Predictores del rendimiento de los alumnos	NO
(Castillejos López, 2022)	Aprendizaje automático Visión por ordenador Robótica Procesamiento del lenguaje natural (PLN) Reconocimiento automático de voz	Actividades contrarias a la ética en la entrega de productos de aprendizaje	NO
(Barrios-Tao et al., 2021)	Realidad aumentada Realidad mixta Regresión logística Bosque aleatorio Minería de datos	Aprendizaje personalizado Aprendizaje adaptativo Inteligencia Instrucción asistida por ordenador Análisis del aprendizaje	NO
(Ayuso del Puerto y Gutiérrez Esteban, 2022)	Chatbots Aprendizaje automático	Herramientas para crear proyectos de aprendizaje automático para la educación en IA	NO
(Jalón Arias et al., 2021)	Aprendizaje automático Procesamiento del lenguaje natural (PLN)	Chatbots web	NO
(Barrios, 2023)	Modelos de lenguaje GPT-3 y GPT-4, aprendizaje supervisado y reforzado	Redacción científica y asistencia en la escritura académica (ChatGPT)	NO
(Hernández Arias, 2023)	Google Bard, ChatGPT; aprendizaje automático y PLN	Apoyo a la investigación científica y generación de contenidos	NO
(Múnera-Duque, 2023)	IA aplicada en simuladores quirúrgicos, robótica educativa	Cirugía educativa asistida por IA (simulación y entrenamiento)	NO

Tabla 2: Herramientas de Inteligencia Artificial aplicadas a la Educación.			
Referencia	Técnicas utilizadas	Aplicación en la educación	Existen pruebas de estrategias de prevención del abandono escolar.
(Lopezosa, 2023)	IA generativa (ChatGPT, Midjourney, Dall-E), PLN y aprendizaje profundo	Comunicación científica con IA generativa	NO
(Forero-Corba y Negre Bennasar, 2024)	Aprendizaje automático: Random Forest, Support Vector Machine, Árboles de decisión, Algoritmos de clustering (K-Means)	Predicción del rendimiento académico y deserción escolar	NO
(García Esquirol, 2015)	IA basada en reconocimiento de lenguaje natural, gamificación y aprendizaje automático	Simulación y diagnóstico médico en la enseñanza clínica (Mediktor)	NO
(Hidalgo Suarez et al., 2023)	Computer-Supported Collaborative Learning (CSCL), IA aplicada en evaluación automática de código y simuladores de docente	Aprendizaje colaborativo asistido por ordenador en programación	NO
(Castillejos López, 2022)	IA en redes sociales (TikTok), PLN, análisis de prácticas educativas poco éticas y reflexión crítica	Evaluación ética del uso de IA en entornos personales de aprendizaje (PLE)	NO
(Jalón Arias et al., 2021)	Análisis PESTEL, Mapas Cognitivos Difusos (MCD), aprendizaje automático	Desarrollo de competencias digitales en educación jurídica con IA	NO
(Vidal Ledo et al., 2019)	Simulación médica avanzada, IA en telemedicina y sistemas expertos	Docencia médica personalizada y simuladores clínicos avanzados	NO
(Gual-Sala, 2023)	Simulación de pacientes virtuales, IA en sistemas expertos para docencia médica	Asistencia en la docencia médica y simulación de pacientes virtuales	NO
(Coto Jiménez, 2021)	Enfoques didácticos en IA, análisis de currículo y estrategias pedagógicas en ingeniería eléctrica	Formación en ingeniería eléctrica incorporando IA en programas de grado	NO
(Forero-Corba y Negre Bennasar, 2024)	- Random Forest (RF) - Árbol de decisión (DT) - K-Nearest Neighbors (KNN) - Redes neuronales convolucionales (CNN) - Redes neuronales artificiales (ANN) - Redes neuronales multicapa feedforward (MFFNN)	- Predicción del rendimiento académico - Detección temprana de deserción - Apoyo en el aprendizaje de estudiantes con TEA - Generación de contenido educativo - Mejora de la orientación académica y profesional	NO
(Bolaño-García y Duarte-Acosta, 2024)	Sistemas de tutoría inteligente - Minería de datos educativos - Modelos computacionales basados en la web semántica	Personalización del aprendizaje - Retroalimentación y evaluación adaptativa - MOOC con soporte de IA	NO
(García Peñalvo et al., 2024)	ChatGPT (Generative Pre-trained Transformer) - Modelos de lenguaje generativo - Modelos de redes neuronales de gran tamaño	Personalización del aprendizaje - Asistentes virtuales para tutoría - Experiencias de aprendizaje inmersivas e interactivas	NO
(Juca-Maldonado, 2023)	ChatGPT (transformador generativo preentrenado)	Generación de contenido académico - Evaluación de la calidad de documentos académicos por IA	NO
(Rodríguez Almazán et al., 2023)	ChatGPT - Aprendizaje profundo basado en redes neuronales "Transformer"	Desarrollo de habilidades en STEM - Retroalimentación personalizada - Aprendizaje interactivo y autónomo	SÍ
(Bustamante Bula y Camacho Bonilla, 2024)	Redes neuronales - Sistemas de predicción basados en datos educativos	Detección de causas de deserción escolar - Mejora del rendimiento académico en secundaria	SÍ

En el ámbito de la educación superior, la predicción de la deserción se relaciona directamente con indicadores de calidad institucional como la tasa de graduación oportuna, la eficiencia terminal y la acreditación de programas. Estudios empíricos han confirmado que la incorporación de modelos de inteligencia artificial mejora significativamente la precisión de los sistemas de alerta temprana en universidades, optimizando la asignación de recursos de apoyo académico y psicosocial (Ibarra-Vazquez et al., 2023; Melchor et al., 2025). De este modo, el uso de herramientas de aprendizaje automático no sustituye, sino que complementa las reformas pedagógicas, al ofrecer evidencia objetiva para focalizar las intervenciones.

Si bien las reformas pedagógicas y curriculares son necesarias para mejorar la pertinencia educativa, estas no garantizan por sí solas la disminución de la deserción, pues suelen atender causas estructurales de largo plazo. La predicción mediante ML, en cambio, permite actuar sobre factores inmediatos y multivariados, como el desempeño académico, la situación socioeconómica o las condiciones emocionales, que impactan directamente la permanencia (Camargo García, 2020; Niyogisubizo et al., 2022). Por ello, la integración de modelos predictivos constituye una estrategia complementaria y sinérgica a las reformas tradicionales.

Se recomienda a las instituciones implementar sistemas de monitoreo académico basados en analítica de datos que, combinados con modelos predictivos, permitan establecer rutas diferenciadas de acompañamiento estudiantil. Estas rutas pueden incluir tutorías personalizadas, apoyos socioeconómicos, ajustes curriculares

flexibles y programas de bienestar institucional, diseñados a partir de la segmentación generada por los algoritmos (Améstica-Rivas et al., 2021; Fuertes Arroyo y Uc Ríos, 2023). De esta manera, se logra articular la evidencia técnica de los modelos con estrategias de política institucional.

La utilización de algoritmos de aprendizaje automático para la detección temprana de factores asociados a la deserción escolar constituye un enfoque novedoso y de gran relevancia en el ámbito educativo contemporáneo. El abandono escolar, considerado un problema persistente y de carácter global, exige estrategias de prevención y gestión eficaces. En este escenario, las técnicas de *Machine Learning* han cobrado protagonismo debido a su capacidad para procesar grandes volúmenes de datos, identificar patrones complejos y anticipar comportamientos de riesgo. Gracias a ello, es posible reconocer con antelación a los estudiantes que presentan mayor probabilidad de abandonar sus estudios, lo que brinda a las instituciones una oportunidad única para intervenir de manera temprana con acciones focalizadas y personalizadas.

No obstante, como se observa en la Tabla 3, la mayoría de los estudios revisados se limita a la construcción y validación de modelos predictivos, sin avanzar hacia la formulación de estrategias de intervención directamente fundamentadas en sus resultados. Aunque la evaluación del desempeño de los algoritmos suele reportarse mediante métricas de precisión, en muchos casos persiste la ausencia de prácticas más robustas, como la validación cruzada sistemática o el ajuste exhaustivo de hiperparámetros. Esta carencia limita la capacidad de trasladar los hallazgos a escenarios aplicados que trasciendan el ejercicio experimental.

Asimismo, se evidencia un uso restringido de métodos de conjunto como *Gradient Boosting*, *Majority Voting* o *Random Forest*, cuyo potencial para mejorar la precisión y estabilidad de las predicciones está documentado en la literatura reciente. Esta limitada adopción puede explicarse por la naturaleza y disponibilidad de los datos educativos utilizados, los cuales no siempre presentan las condiciones idóneas para la implementación de modelos más avanzados. Finalmente, ninguno de los trabajos analizados integra estos enfoques predictivos en el marco de programas de formación basados en competencias, lo que resalta una oportunidad de investigación aún no explorada. Dado que la educación por competencias constituye una tendencia creciente en la educación superior, resulta pertinente profundizar en cómo estos modelos podrían articularse con políticas institucionales orientadas a la permanencia estudiantil.

De esta manera, la Tabla 3 sintetiza la aplicación de diferentes algoritmos de aprendizaje automático en el contexto educativo, pero deja en evidencia la falta de articulación entre los resultados predictivos y el diseño de estrategias de intervención efectivas, brecha que este estudio busca contribuir a cerrar.

Tabla 3: Modelos estadísticos de Machine Learning en la previsión de la deserción y la mejora del rendimiento de los estudiantes.

Referencia	Algoritmos utilizados
(Viloria et al., 2019)	Naive Bayes BayesNet
(Valero Cajahuanca et al., 2022)	SVM Regresión logística Árbol de decisión KNN
(Treviño et al., 2013)	Lógica difusa
(Tete et al., 2022)	Árboles de decisión Redes bayesianas Regresión logística Redes neuronales Máquinas de vectores soporte Reglamento de la Asociación
(Smith Uldall y Gutiérrez Rojas, 2022)	Regresión logística Árboles de decisión Bosque aleatorio Redes neuronales
(Andrés Rico, 2022)	Bayes ingenuo K Vecinos más próximos Árbol de decisión C4.5
(Pineda-Pertuz et al., 2022)	Árboles de decisión Bosque aleatorio Refuerzo adaptativo Refuerzo de gradiente Regresión logística Máquinas de vectores soporte

Tabla 3: Modelos estadísticos de Machine Learning en la previsión de la deserción y la mejora del rendimiento de los estudiantes.

Referencia	Algoritmos utilizados
(Jimenez Chaves y García Torres, 2019)	K - Medios Análisis de componentes principales (ACP)
(Hoyos Osorio y Daza Santacoloma, 2023)	Regresión logística
(Guzmán-Castillo et al., 2022)	AdaBoost Bayesiano GLM Árboles de decisión Logit Boost Random Forest Gradiente estocástico Boosting.
(Gutierrez-Pachas et al., 2023)	Regresión logística Máquinas de vectores soporte Bayas ingenuas gaussianas Vecinos más próximos (KNN) Árboles de decisión Bosque aleatorio Perceptrón multicapa (MP) Red neuronal convolucional (CNN) - Relu
(Flores et al., 2022)	Bosque aleatorio Árbol aleatorio J48 REPTree JRIP OneR Red Bayes Bayes ingenuo
(Fernández-Martín et al., 2019)	Regresión multinomial Bosque aleatorio KNN Árboles potenciados Árboles de decisión Redes neuronales Máquinas de vectores soporte (SVM) Regresión logística
(Contreras Bravo et al., 2022)	KNN VPC Bayes ingenuo Análisis discriminante lineal (LDA) Árboles de decisión
(Cardozo et al., 2022)	Regresión logística Redes bayesianas Árboles de decisión
(Bressane et al., 2022)	Análisis discriminante lineal Regresión logística Red de Correlación en Cascada (CCN) Impulso para los árboles Sistema de inferencia difusa (FIS) Programación de la expresión génica (GEP) Red Perceptrón Multicapa (MLP) Red neuronal polinómica (GMDH) Red neuronal probabilística (PNN) Red de función de base radial (RBFN) Máquinas de vectores soporte
(Bitencourt et al., 2022)	Máquinas de vectores soporte Refuerzo de gradiente
(Albán y Mauricio, 2018)	Árboles de decisión
(Al Ka'bi, 2023)	Redes neuronales convolucionales (CNN)
(Henriquez Cabezas y Vargas Escobar, 2022)	Modelo logístico multivariante
(Incio-Flores et al., 2023)	Red neuronal Levenberg - Marquardt Red neuronal Gradiente conjugado escalado
(Gil-Vera y Quintero-López, 2021)	Bayes ingenuo Regresión logística Máquinas de vectores soporte KNN Bosque aleatorio Árboles de decisión
(Castrillón et al., 2020)	Algoritmo bayesiano de clasificación J48
(Christou et al., 2023)	RBFNN, DT, BN, SVM, NNC, LM-BP, BFGS-BP, SGD
(Ibarra-Vazquez et al., 2023)	Árboles de decisión, bosques aleatorios

Tabla 3: Modelos estadísticos de Machine Learning en la previsión de la deserción y la mejora del rendimiento de los estudiantes.	
Referencia	**Algoritmos utilizados**
(Niyogisubizo et al., 2022)	Random Forest (RF), XGBoost, Gradient Boosting (GB), FNN
(Gonzalez Salas Duhne et al., 2022)	Regresión logística, Chi-cuadrado, Aprendizaje automático supervisado
(Delogu et al., 2024)	Bosques aleatorios, máquinas de aumento gradual (GBM)
(Krüger et al., 2023)	Árboles de decisión, clasificadores conjuntos, clasificadores monolíticos
(Kordbagheri et al., 2025)	kNN, Gradient Boosted Ensemble (GBE), Random Forest (RF)
(Martínez y Castillo, 2024)	Árbol potenciado por gradiente (GBT)
(Castro-Maldonado et al., 2022)	K-Means, PCA
(Londoño-Gallego et al., 2024)	DBSCAN, Dendograma, K-Means, PCA
(Rodríguez et al., 2023)	Árbol de decisión XGBoost LuzGBM CatBoost
(Ali et al., 2025)	Máquina de vectores soporte Lineal múltiple Cresta Red elástica LASSO
(Mustofa et al., 2025)	Bayes ingenuo Bosque aleatorio Máquina de vectores soporte Regresión logística XGBoost Red neuronal artificial (RNA) Perceptrón multicapa (MLP) Regresión logística híbrida y red neuronal (HLRNN)
(Rabelo y Zárate, 2025)	Log R ANN CART discretizado ANN Bosque aleatorio Ensemble - sistemas de votación (CART, Log R y ANN)

Fuente: Autor

En este contexto, es esencial abordar la deserción escolar de manera efectiva y proactiva, desarrollando estrategias de prevención y retención que puedan identificar y apoyar a los estudiantes en riesgo de abandonar la educación, es así que diferentes autores ya se encuentran abordando este tema reflexionando sobre la importancia de la implementación de nuevas tecnologías encaminadas al uso de algoritmos de inteligencia artificial para identificar modelos que permitan encontrar patrones dentro de la minería de datos educativa y realizar pronósticos sobre la deserción o el desempeño académico de los estudiantes.

Los modelos que se vienen identificando y aplicando en los procesos educativos van desde temas relacionados con la identificación de patrones ocultos orientados a definir las fortalezas en la implementación de ciertas estrategias y técnicas didácticas hasta Sistemas de tutoría inteligente (STI) que permiten optimizar la experiencia de aprendizaje de los estudiantes. Respecto al fenómeno de la deserción, se vienen usando modelos estadísticos de aprendizaje automático para reconocer o identificar patrones ocultos a partir del Big Data académico de las instituciones y, predicciones del comportamiento académico de los estudiantes con base a información multidimensionalidad de las calificaciones de las materias e información registrada al momento de la matrícula. Para la implementación de modelos de aprendizaje automático, es fundamental contar con un marco estructurado para su diseño y construcción, a fin de obtener el mejor modelo que se adapte al contexto.

En este sentido, el presente trabajo se destaca al integrar modelos de aprendizaje automático supervisados y no supervisados para prevenir y contrarrestar la deserción estudiantil en una institución de formación por competencias en Colombia. En coherencia con lo anterior, el objetivo principal del artículo es demostrar la viabilidad y eficacia de un modelo de aprendizaje automático para predecir la deserción estudiantil en una institución colombiana de formación por competencias. De manera específica, se busca: (1) identificar variables críticas relacionadas con el riesgo de abandono; (2) comparar el desempeño de diferentes algoritmos; y (3) diseñar un marco de acción institucional basado en la segmentación estudiantil.

2. Materiales y Método

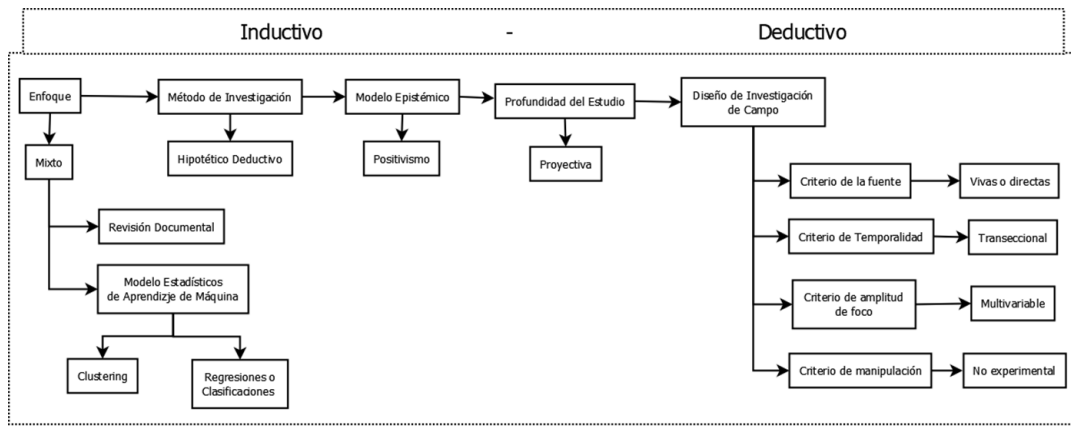
2.1. Metodología de la investigación

En términos metodológicos, este estudio responde a una brecha de investigación identificada en la literatura: aunque existen múltiples trabajos sobre predicción de la deserción, pocos abordan su aplicación en instituciones de educación por competencias, un contexto particularmente relevante en Colombia y aún poco explorado. La contribución novedosa de esta investigación radica en el desarrollo de un modelo híbrido que combina algoritmos supervisados y no supervisados, lo que permite no solo predecir el riesgo de abandono, sino también segmentar a los estudiantes en perfiles diferenciados que facilitan la definición de políticas institucionales de intervención. La elección de este tema se justifica porque, a diferencia de enfoques pedagógicos tradicionales o análisis meramente descriptivos, el uso de *Machine Learning* ofrece una herramienta empírica y proactiva para la gestión de la permanencia estudiantil. Adicionalmente, la propuesta permite estructurar un modelo operativo formal que incorpora técnicas de *Machine Learning* con el fin de diseñar y aplicar estrategias de mitigación de la deserción escolar, lo que incrementa su relevancia práctica. De esta manera, el artículo realiza una aportación interdisciplinaria que conecta tres dominios complementarios: la educación/pedagogía, al enfocarse en la retención y en la mejora de la calidad formativa; los medios de información y la analítica de datos, al emplear técnicas de minería y modelado predictivo; y la inteligencia artificial, al validar algoritmos de aprendizaje automático en un entorno real. En consecuencia, la significación de este trabajo se centra en demostrar cómo la integración de la IA en instituciones de educación superior basadas en competencias puede convertirse en una estrategia innovadora y robusta para enfrentar la deserción estudiantil.

De acuerdo con Hurtado de Barrera (2012), este trabajo tendrá un enfoque mixto, toda vez, que inicialmente se explorarán y analizarán documentos sobre los métodos, metodologías y modelos de aprendizaje automático aplicados a la predicción del fenómeno de la deserción escolar en diferentes contextos. Sobre esta búsqueda serán identificados los mejores modelos, que serán aplicados al contexto de los datos de estudio para encontrar patrones ocultos o pronosticar el comportamiento del fenómeno a partir de las matrices de características. Igualmente, la hipótesis que presenta este trabajo establece que los modelos estadísticos de aprendizaje automático pueden pronosticar y caracterizar la decisión de los estudiantes de abandonar o no sus estudios en instituciones de formación por competencias. Este enfoque se alinea con un paradigma positivista, ya que busca confirmar la mencionada hipótesis.

De otra parte, el diseño de investigación se enmarca según el criterio de la fuente en vivas o directas; según la temporalidad en transaccional, lo que implica la recopilación de datos en momentos específicos; según amplitud de foco en multivariable, por tener en cuenta múltiples variables en el análisis; y según el criterio de manipulación se clasifica como no experimental. Finalmente, el presente estudio se plantea como una investigación de carácter proyectivo, toda vez, que plantea una metodología tanto para la aplicación de algoritmos de *Machine Learning* en el contexto específico como en el desarrollo de un modelo operativo formal. Así, este trabajo no solo se limita a identificar las características del objeto de estudio, sino que propone una metodología innovadora para integrar herramientas de inteligencia artificial en la prevención de la deserción escolar. (Figura 2)

Figura 2: Metodología de la investigación.



2.2. Método operativo de la investigación

2.2.1. Caracterización de los datos

Los datos fueron recopilados de diversas fuentes. En primer lugar, se solicitaron los datos de caracterización sociodemográfica al administrador de la plataforma LMS utilizada por la entidad para el registro académico de los estudiantes. Para el seguimiento del rendimiento académico se diseñó un micrositio específicamente para este estudio donde se registraron las quejas del comportamiento académico de los estudiantes, las solicitudes de cambios de calificaciones realizadas por los docentes en respuesta a los diferentes refuerzos pedagógicos y los planes de mejora asignados a los estudiantes (ver Tabla 4)

Tabla 4: Bases de datos usadas para la investigación.

Dimensión	Atributo	Tipo	Tamaño (filas y columnas)
Sociodemográfica	Departamento Municipio Género Edad Estrato Estado Civil	Categórico Categórico Categórico Categórico Categórico Categórico	1997 x 9
Seguimiento Académico	Cantidad de quejas Cantidad de reversiones	Numérico Numérico	2945 x 12 800 x 19
Desempeño Académico	Resultados de aprendizaje	Categórico	De acuerdo con la base de datos del programa

Finalmente, los datos sobre el desempeño académico se obtuvieron de la misma plataforma LMS de la entidad, pero accediendo desde el rol de coordinación académica. El tamaño del conjunto de datos destinado a evaluar el desempeño académico de los estudiantes varió según las bases de datos de cada uno de los programas, tal como se observa en la Tabla 5.

Tabla 5: Bases de datos de los programas de formación.

Programa	Filas	Columnas	Atributos	Tipo	Clase
Animación 3D	32	75	Resultados de aprendizaje de las competencias del programa de formación	Categórico (Aprobado, No aprobado, Por evaluar)	Traslado En formación Condicionado Cancelado Retiro voluntario
Gestión de Redes de Datos	236	82			
Impresión Digital	49	72			
Preprensa	91	80			
Producción de Medios Audiovisuales	116	86			
Producción Multimedia	155	70			
Sistemas	210	66			
Desarrollo de Medios Gráficos Visuales	49	78			
Videojuegos	53	76			
Elaboración de Audiovisuales	48	64			
Mantenimiento de equipos de cómputo	20	64			
Análisis y Desarrollo de Software	640	85			
Animación Digital	29	78			

2.2.2. Minería de datos

Se consideró la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) cuyo método proporciona una estructura organizada para el desarrollo de actividades relacionadas con la ingeniería y la ciencia de datos. La metodología CRISP-DM se utilizó para establecer un workflow relacionado a la adquisición, evaluación y estructura de los datos y, contribución significativa o no de los resultados obtenidos. Las etapas que contempla la metodología CRISP-DM involucra el entendimiento, preparación y modelado de los datos; y posterior evaluación y despliegue del modelo. Para la etapa de entendimiento y preparación fueron creadas las bases de datos correspondientes a las condiciones sociodemográfica, seguimiento y desempeño académicos (ver Tabla 4), obtenidos a partir de la implementación de un micrositio (*website* de Google) dispuesto por la coordinación académica (desde junio 2022 hasta octubre 2023), que fueron articuladas con las bases de datos de los programas académicos bajo estudio (ver Tabla 2), tomados de las herramientas LMS (Learning Management Sistem) de la entidad.

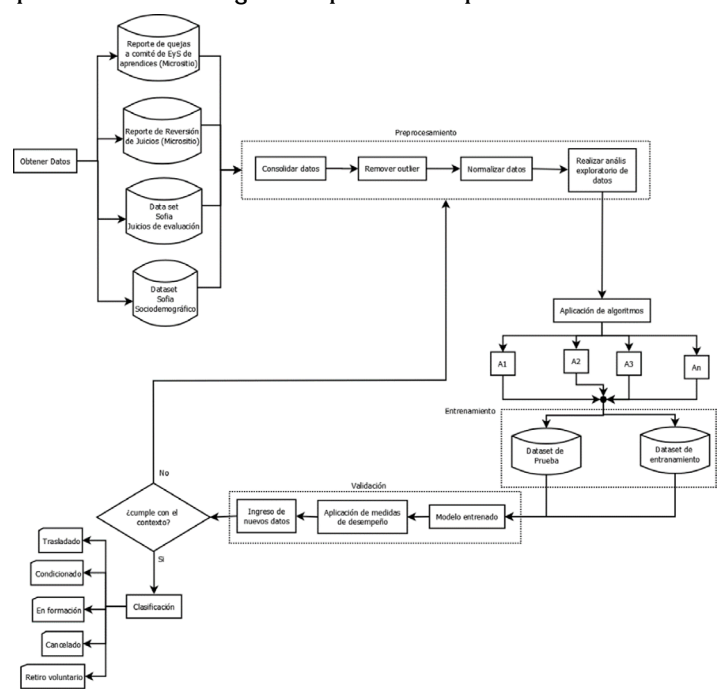
Durante la etapa de modelado, se llevó a cabo el entrenamiento, las pruebas y la validación de una variedad de algoritmos documentados en la literatura científica. La elección de los modelos se realizó basándose en el criterio de benchmarking con el objetivo de identificar aquellos que han demostrado ser

exitosos en contextos similares. Asimismo, se aplicó el criterio de experimentación para implementar modelos innovadores que, según la opinión de expertos, podrían arrojar resultados prometedores en términos de métricas de rendimiento. En la fase subsiguiente, se procedió a la calibración de los hiperparámetros utilizando métodos como la validación *holdout* y *k-fold cross-validation*. Esto se hace con el propósito de determinar los valores óptimos que maximicen la eficacia del modelo. Para llevar a cabo las actividades de consolidación, arreglo, preprocesamiento y procesamiento de los datos y modelado del algoritmo, se empleó el entorno de Google Colab® utilizando el lenguaje de programación Python.

En la Figura 3 se representa el flujo de procesamiento y validación de un modelo de aprendizaje automático para el caso específico. El proceso comenzó con la obtención de datos de diferentes fuentes, que incluyen reportes de quejas, reversión de juicios y datasets de evaluación y datos sociodemográficos. Estas fuentes proporcionan información diversa y detallada sobre el contexto académico de los estudiantes, lo que permite identificar factores potenciales asociados a la deserción. Una vez obtenidos los datos, se inició la fase de preprocesamiento, que incluye la consolidación de los datos, la eliminación de valores atípicos (outliers) que puedan afectar el análisis, y la normalización para garantizar que todas las variables estén en una escala comparable. Posteriormente, se realizó un análisis exploratorio de los datos, con el objetivo de identificar patrones iniciales y seleccionar las variables más relevantes para el modelo.

Tras el preprocesamiento, los datos se dividieron en conjuntos de entrenamiento y prueba, y se procede con la aplicación de algoritmos de aprendizaje automático (A1, A2, A3, ... An). Estos algoritmos se entrenaron para identificar patrones en los datos que puedan predecir la deserción. Después del entrenamiento, se realizó la fase de validación, donde se ingresan nuevos datos para probar el modelo, aplicando métricas de desempeño que permitan evaluar su precisión y generalización. Finalmente, el modelo se sometió a un proceso de clasificación, donde, según el contexto de los estudiantes (traslado, condicional, en formación, cancelado, o retiro voluntario), se determina su estado. Este proceso de validación y clasificación aseguró que el modelo no solo sea preciso en términos de predicción, sino también relevante y aplicable en el entorno educativo donde se implementará. Al culminar este proceso, se obtuvo un modelo o modelos de datos entrenados y validados con hiperparámetros definidos. Este modelo o modelos se utilizarán para clasificar a los estudiantes en categorías relacionadas con el dataset, tales como "Traslado", "Condicionado", "En formación", "Cancelado" y "Retiro voluntario". (Figura 3)

Figura 3: Método operativo de la investigación a partir de la aplicación de modelos supervisados de ML.

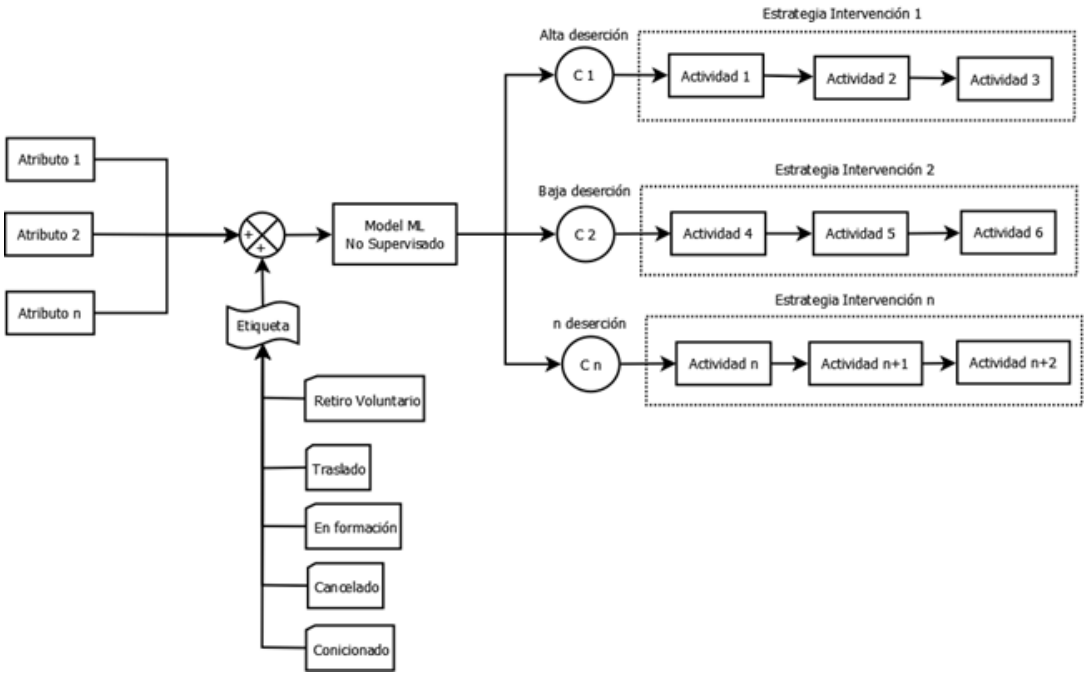


De otro lado, la Figura 4 presenta un flujo de análisis y segmentación de patrones de deserción utilizando un modelo de aprendizaje automático no supervisado. Este modelo permitió agrupar estudiantes en diferentes categorías de deserción según una serie de atributos relevantes (Atributo 1, Atributo 2, ..., Atributo n) que pueden incluir características demográficas, académicas, de comportamiento o incluso algunas características (o etiquetas) ya definidas por las bases de datos, por tanto, se amplía el conjunto de datos adicionando la etiqueta presentada por el proceso de formación por competencias de la institución ("Traslado", "Condicionado", "En formación", "aplazado", "Cancelado" y "Retiro voluntario") como atributo adicional de los sujetos de estudio. A diferencia de los modelos supervisados, en los modelos no supervisados no se cuenta con una "etiqueta" predefinida para cada caso; en cambio, el modelo identificó patrones o grupos (clusters) de deserción de manera autónoma, basándose únicamente en las relaciones entre los atributos.

Cada grupo identificado por el modelo (C1, C2, ..., Cn) representa un tipo específico de deserción con características comunes, como alta, media y baja deserción.

Este procedimiento orientado a combinar modelos no supervisados con modelos estadísticos supervisados ya se viene desarrollando por otros autores, en donde se ha vislumbrado una potente estrategia para ajustar las predicciones de los algoritmos con bases a un dataset con gran cantidad de etiquetas (Parkavi et al., 2023).

Figura 4: Clústeres determinados por el modelo no supervisado de ML.



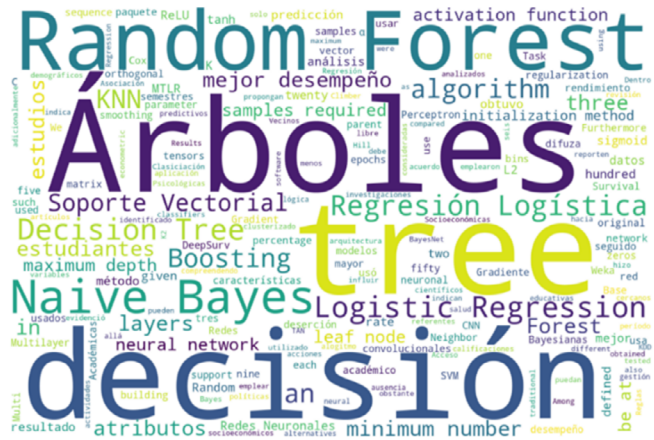
Asimismo, se emplearon métricas de rendimiento específicas, como precisión, sensibilidad, especificidad y F1-score, que permitieron evaluar no solo la efectividad del modelo para identificar patrones, sino también su equilibrio en la detección de casos positivos y negativos. Estos análisis contribuyen a determinar si el modelo es realmente eficaz y aplicable en escenarios reales donde los datos pueden variar en función de la institución, la región o la población estudiantil. Una vez identificadas y estructuradas las características más relevantes de cada uno de los clústeres, se propondrá la formulación de estrategias de intervención específicas, generadas mediante herramientas de inteligencia artificial generativa, para cada uno de los grupos identificados.

3. Resultados y discusión

La aplicación de algoritmos de inteligencia artificial ha traspasado fronteras de saberes y conocimientos, incluso para optimizar los procesos de investigación cualitativa, en ese sentido, se optó por usar la librería

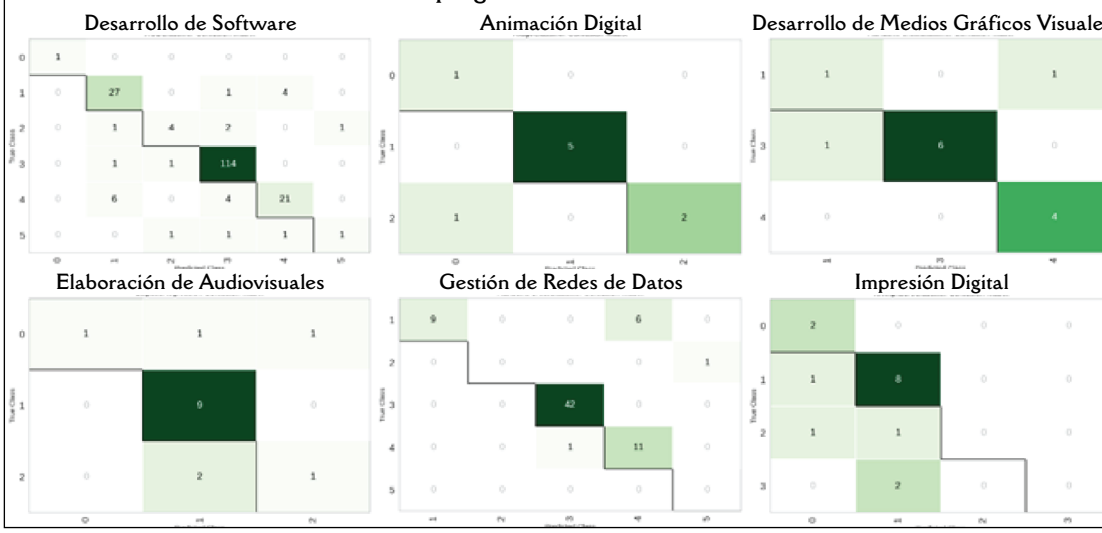
de procesamiento de lenguaje natural (PLN) como es Natural Language Toolkit (NLTK) donde pudimos generar una nube de palabras (ver Figura 5), en ella se puede observar la frecuencia de uso de los diferentes algoritmos de Machine Learning en el contexto educativo. Podemos inferir que el algoritmo más usado en estas aplicaciones, de acuerdo con los documentos consultados, es árbol de decisión seguido del Random Forest y, por último, *Naive Bayes*, esto en coherencia a lo que indicó Tete et al. (2022) en donde relaciona estos mismos algoritmos como resultados de su revisión sistemática realizada entre los años 2010 a 2020.

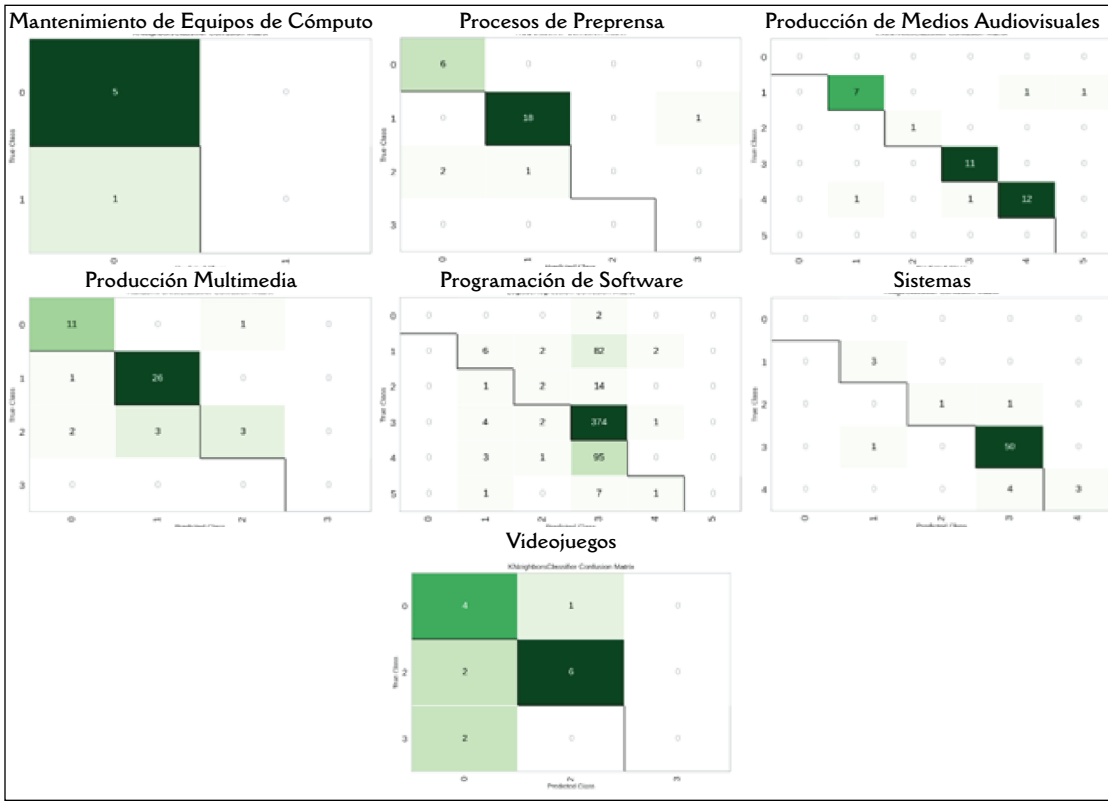
Figura 5: Nube de palabras con los algoritmos más comunes encontrados de la literatura consultada.



Con base en las bases de datos recopiladas y siguiendo las etapas definidas en la metodología CRISP-DM, se llevó a cabo la exploración de distintos modelos de aprendizaje automático supervisados aplicables al contexto de la deserción estudiantil. Para esta fase, se utilizó la biblioteca de código abierto PyCaret, la cual facilita la automatización de flujos de trabajo en *Machine Learning*. La herramienta fue aplicada a cada uno de los programas analizados, empleando como criterio de comparación de desempeño la matriz de confusión. En ese sentido, en la Figura 6 se presentan las matrices de confusión de los modelos con mejor desempeño en cada uno de los programas de formación estudiados, teniendo en cuenta las clases o etiquetas iniciales ("Traslado", "Condicionado", "En formación", "Aplazado", "Cancelado" y "Retiro voluntario").

Figura 6: Matriz de confusión de los modelos con los mejores desempeños de cada uno de los programas estudiados.





Asimismo, la Tabla 6 presenta un resumen de los algoritmos de clasificación con mejor desempeño, así como la frecuencia con la que fueron seleccionados como los más efectivos en los distintos programas de formación analizados. Se observa que el *Random Forest Classifier* y el *K Neighbors Classifier* fueron los algoritmos más frecuentemente seleccionados, con una aparición destacada en tres ocasiones cada uno, lo que sugiere una alta robustez y capacidad de generalización en el contexto del problema de predicción de deserción estudiantil. Le siguen el *Ridge Classifier* con dos apariciones, y otros algoritmos como *XGB Classifier*, y *Logistic Regression Classifier* cada uno seleccionado dos veces. Esta distribución evidencia la predominancia de modelos basados en ensamblado y regularización, lo cual es consistente con la naturaleza compleja y posiblemente no lineal de los datos analizados.

Tabla 6: Algoritmos con mejor desempeño con Dataframe Completo.

Algoritmos	Frecuencia
RandomForestClassifier	3
KNeighborsClassifier	3
RidgeClassifier	2
XGBClassifier	2
LogisticRegression	2
ExtraTreesClassifier	1

Estos hallazgos son consistentes con estudios previos. Por ejemplo, (Camargo García, 2020), en una investigación realizada en la Universidad de la Costa (Colombia), reportaron que el algoritmo *Random Forest* logró una precisión del 84.8% al predecir la deserción estudiantil, superando a modelos como las redes bayesianas y las máquinas de vectores de soporte. Este rendimiento se atribuye a la capacidad del algoritmo para manejar datos mixtos y capturar relaciones complejas entre variables sin necesidad de un preprocesamiento exhaustivo. De igual forma, estos resultados son coherentes con hallazgos de otros estudios, en los cuales el algoritmo *Random Forest Classifier* ha demostrado un desempeño destacado en contextos similares, superando o igualando en eficacia a otros modelos de clasificación (Ibarra-Vazquez et al., 2023; Martínez y Castillo, 2024; Niyogisubizo et al., 2022; Tete et al., 2022).

Con el propósito de que los hallazgos de esta investigación puedan integrarse de manera efectiva en los procesos institucionales orientados a la mitigación del abandono académico, y sean aplicables mediante una herramienta tecnológica, se procedió a priorizar la identificación de las variables más relevantes para la predicción del estado de los estudiantes. En este sentido, se seleccionaron las siguientes características del *dataframe* procesado: "Cantidad de quejas", "Cantidad de reversiones", "Género", "Edad", "Estado civil" y "Estrato", con el objetivo de predecir la variable dependiente "Estado Aprendiz".

La Tabla 7 presenta un resumen de los algoritmos que alcanzaron el mejor rendimiento al aplicarse sobre los subconjuntos priorizados de datos correspondientes a los distintos programas de formación. Se evidencia que los algoritmos *KNeighborsClassifier* y LGBMClassifier obtuvieron los mejores desempeños; sin embargo, estos resultados no coinciden con los algoritmos que presentaron el mejor rendimiento al utilizar el *dataframe* completo, lo que sugiere diferencias en la capacidad predictiva según el conjunto de variables analizado, No obstante, la priorización de variables, desde una perspectiva orientada al entendimiento del negocio, resulta necesaria para facilitar la interacción del usuario con la herramienta (aplicativo). Esto se debe a que no sería óptimo requerir el ingreso de un número excesivo de datos para realizar la predicción de la variable objetivo, ya que podría afectar la usabilidad y eficiencia de la herramienta.

Tabla 7: Algoritmos con mejor desempeño con Dataframe priorizado.

Algoritmos	Frecuencia
KNeighborsClassifier	3
LGBMClassifier	3
LogisticRegression	2
RidgeClassifier	2
Logistic Regression	1
AdaBoostClassifier	1
LogisticRegression	1

Fuente: Autor.

No obstante, es importante considerar que la efectividad de los algoritmos de aprendizaje automático puede variar según el contexto y las características del conjunto de datos. Esto sugiere que, si bien *Random Forest* es una opción robusta, la selección del algoritmo óptimo debe basarse en un análisis contextualizado de los datos y del problema educativo abordado.

El algoritmo *Random Forest*, introducido por (Breiman, 2001), es un método de aprendizaje supervisado basado en el principio de ensamblado de múltiples árboles de decisión. Su fortaleza radica en combinar la aleatoriedad en la construcción de árboles con técnicas de agregación para obtener un modelo robusto, con baja varianza y alta capacidad de generalización.

Cada árbol del bosque se construye utilizando una muestra aleatoria (con reemplazo) del conjunto de datos original, conocida como muestra *bootstrap*. Además, en cada división de nodo se selecciona aleatoriamente un subconjunto de características para determinar la mejor partición, reduciendo así la correlación entre los árboles y promoviendo la diversidad del modelo.

Durante la construcción del árbol, se busca encontrar, en cada nodo, la división que minimice una función de impureza. Para tareas de clasificación, dos medidas comunes son el índice de Gini y la entropía. El índice de Gini en un nodo t se define como:

$$G(t) = 1 - \sum_{K=1}^{K} p_{k|t}^2$$

Donde $p_{(k|t)}$ representa la proporción de muestras de clase k en el nodo t, y K es el número total de clases. La entropía se calcula como:

$$H(t) = -\sum_{k=1}^{K} p_{k|t} log_2(P_{k|t})$$

Ambas métricas cuantifican el grado de mezcla de clases en un nodo. El objetivo del algoritmo es seleccionar la división que produzca nodos hijos con menor impureza total.

Para tareas de regresión, en lugar de impureza categórica se emplea la minimización del error cuadrático medio (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Donde y_i es el valor real y $\hat{y}_i$ es la predicción del modelo para cada observación.
Una vez construidos B árboles independientes $\{h_i(x), h_l(x), \ldots, h_B(x)\}$, el modelo Random Forest realiza una agregación para producir su predicción final. Esta agregación difiere según el tipo de problema:
En clasificación, se utiliza votación mayoritaria:

$$\hat{y} = mode\{h_b(x)\}_{b=1}^{B}$$

En regresión, se calcula el promedio de las salidas individuales:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} h_b(x)$$

Esta estrategia de combinación permite reducir la varianza del modelo sin aumentar significativamente el sesgo, una propiedad derivada del principio de agregación de modelos débiles con baja correlación.

Continuando con el desarrollo del proceso investigativo orientado a la identificación de patrones ocultos mediante la aplicación de algoritmos no supervisados, como K-Means y DBSCAN, se implementó inicialmente un Análisis de Componentes Principales (PCA). Esta técnica permitió la visualización del círculo de correlaciones, facilitando la identificación de relaciones entre los distintos atributos del conjunto de datos, incluyendo la etiqueta original. A partir de los eigenvalores y eigenvectores obtenidos, fue posible segmentar los aprendices en clústeres significativos. En la Figura 7 se presenta el círculo de correlaciones junto con el diagrama de dispersión correspondiente a cada uno de los programas analizados.

Con el objetivo de seleccionar el modelo con mejor desempeño, se procedió al análisis de las métricas de validación interna Silhouette Score y Calinski-Harabasz Score. Los resultados obtenidos evidenciaron que, en la mayoría de los programas analizados, el algoritmo K-Means con tres clústeres presentó el mejor rendimiento, permitiendo una segmentación coherente y representativa de los aprendices dentro del contexto del estudio. Teniendo en cuenta estos resultados, los clústeres reportados en el análisis corresponden a los generados por dicho algoritmo.

El algoritmo K-means es un método de aprendizaje no supervisado utilizado para la agrupación (clustering) de datos. Su objetivo es particionar un conjunto de observaciones $X = \{x_1, x_2, \ldots, x_n\}$ con $x_1 \in \mathbb{R}^d$, *en K* grupos (clusters) disjuntos $C = \{c_1, c_2, \ldots, c_k\}$, de modo que la variabilidad intracluster se minimice.

El algoritmo busca minimizar la suma de distancias cuadradas dentro de los clústeres (within-cluster sum of squares, WCSS), formalmente expresada como:

$$\frac{min}{c_1,\ldots,c_K}\sum_{k=1}^{K}\sum_{x_i \epsilon C_k}\|x_i - \mu_k\|^2$$

Donde:

$$\mu_k = \frac{1}{|C_k|}\sum_{x_i \epsilon C_k} x_i \; es\ el\ centroide\ del\ clúster\ C_k,$$

$\|.\|$ denota la norma euclediana en $\mathbb{R}^d$,
$|C_k|$ Es la cantidad de puntos asignados al clúster k
El algoritmo K-means emplea una estrategia iterativa de optimización basada en dos pasos:
Asignación de Clúster (Paso E, Expectation)
Para cada punto x_i, se asigna al clúster más cercano según la distancia euclidiana al centroide actual:

$$c_i = arg\ min_k\|x_i - \mu_k\|^2$$

Actualización de Centroides (Paso M, Maximization)
Se recalcula el centroide de cada clúster como el promedio de los puntos asignados:

$$\mu_k = \frac{1}{|C_k|}\sum_{x_i \epsilon C_k} x_i$$

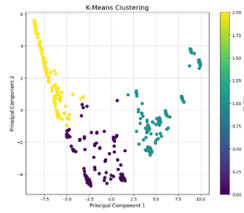
Estos pasos se repiten hasta que las asignaciones de clúster no cambian entre iteraciones (convergencia), o se alcanza un número máximo de iteraciones predefinido. El algoritmo converge a una solución local

óptima del criterio de suma de cuadrados dentro del clúster WCSS (*Within-Cluster Sum of Squares*), aunque no necesariamente al óptimo global, por lo que se recomienda ejecutarlo varias veces con inicializaciones distintas. (Figura 7)

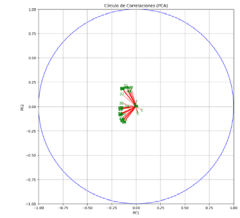
Figura 7: círculo de correlaciones junto con el diagrama de dispersión correspondiente a cada uno de los programas analizados.

Desarrollo de Software

Diagrama de Dispersión

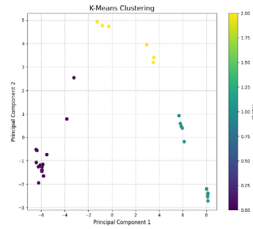


Círculo de Correlaciones

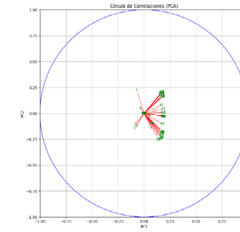


Animación 3D

Diagrama de Dispersión

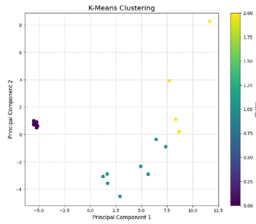


Círculo de Correlaciones

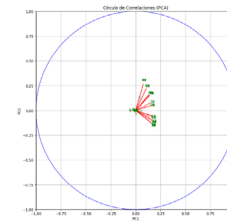


Animación Digital

Diagrama de Dispersión

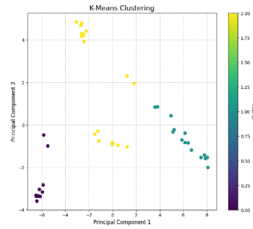


Círculo de Correlaciones

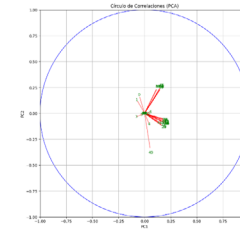


Desarrollo de Medios Gráficos Visuales

Diagrama de Dispersión

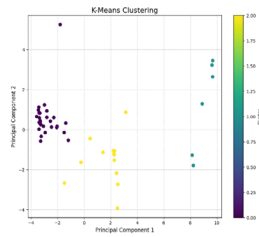


Circulo de Correlaciones

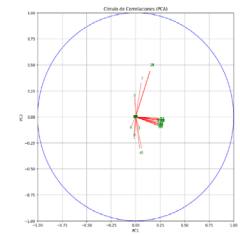


Elaboración de Audiovisuales

Diagrama de Dispersión

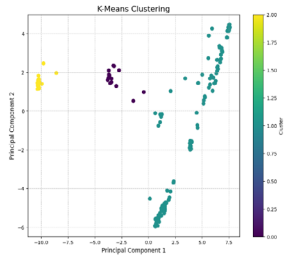


Circulo de Correlaciones

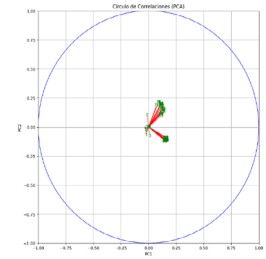


Gestión de Redes de Datos

Diagrama de Dispersión

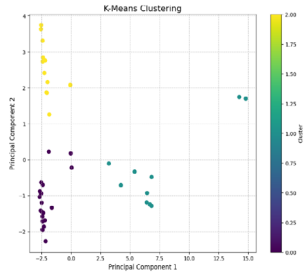


Círculo de Correlaciones

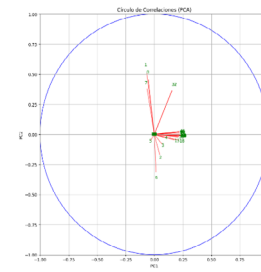


Impresión Digital

Diagrama de Dispersión

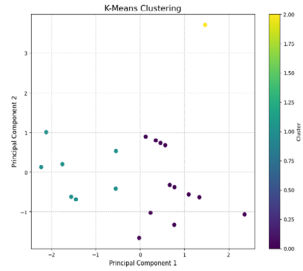


Círculo de Correlaciones

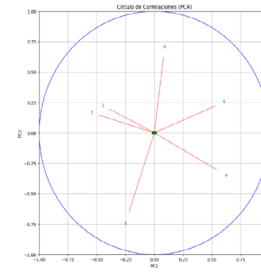


Mantenimiento de equipos de cómputo

Diagrama de Dispersión

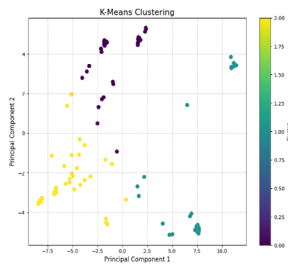


Círculo de Correlaciones

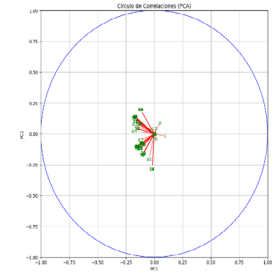


Medios Audiovisuales

Diagrama de Dispersión

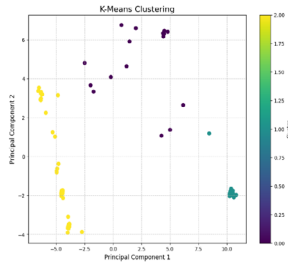


Círculo de Correlaciones

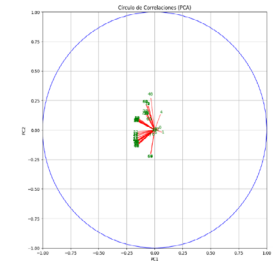


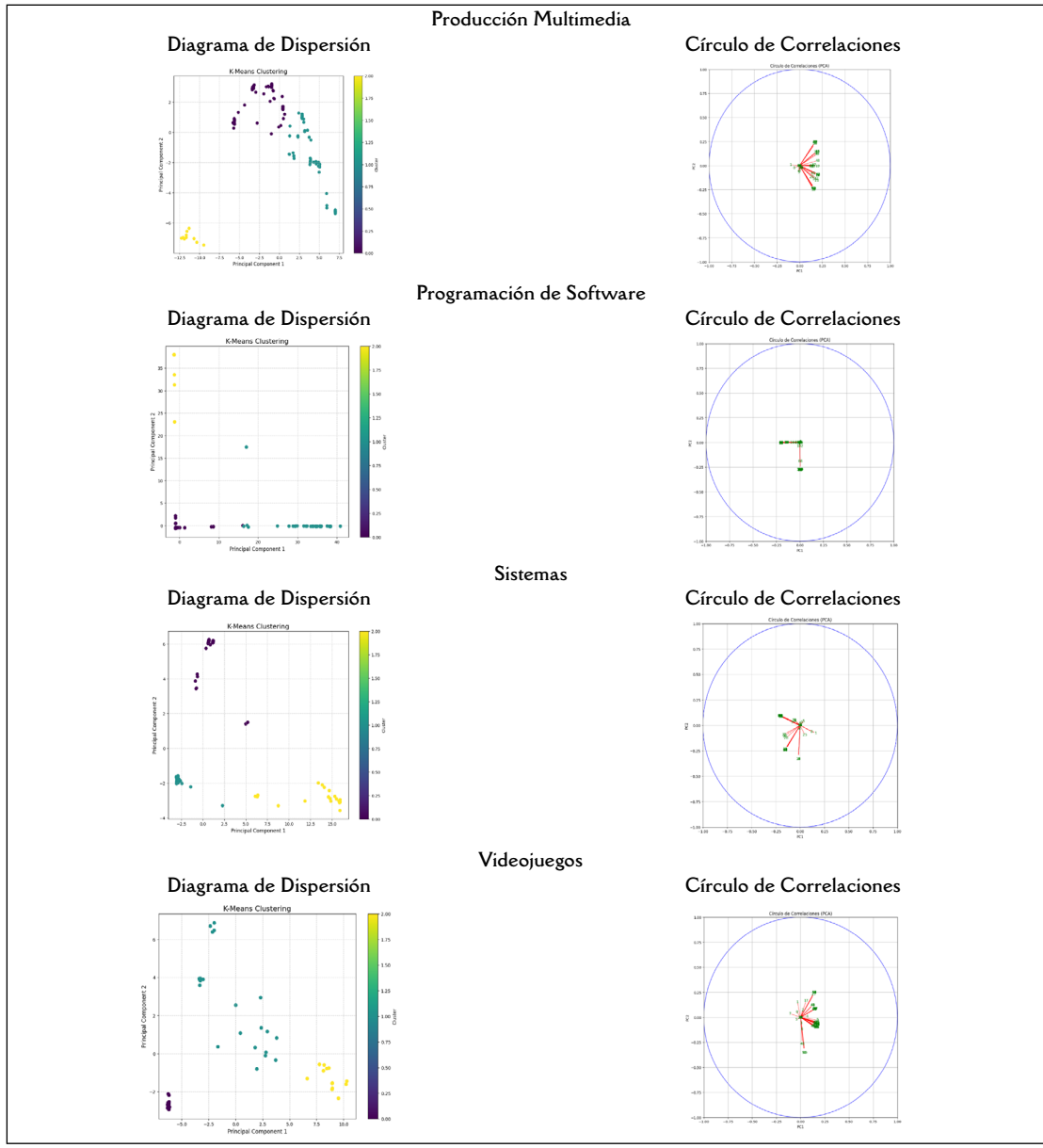
Preprensa

Diagrama de Dispersión



Círculo de Correlaciones



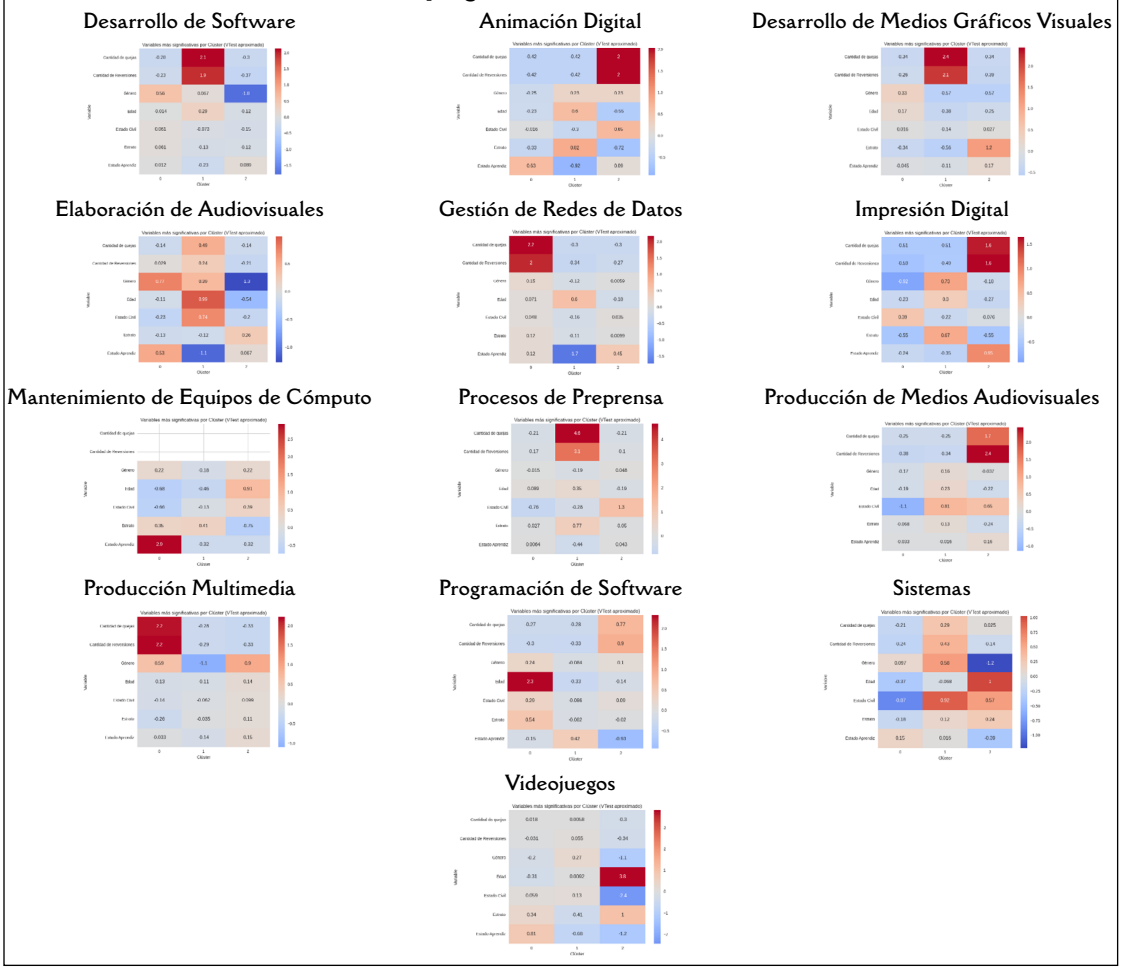


A partir de los clústeres identificados y de la correlación de los vectores de características obtenida mediante el Análisis de Componentes Principales (PCA), se propone un proceso de segmentación orientado a estimar el nivel probable de deserción de los aprendices, categorizado en tres niveles: alto, medio y bajo.

Para contrastar los resultados obtenidos del Análisis de Componentes Principales (PCA) y de los vectores de correlación, se aplicó la metodología de Detección de Clases Latentes (DECLA), integrando la medida estadística *v*-test con el propósito de identificar las variables con mayor capacidad explicativa dentro de los clústeres generados (Magidson y Vermunt, 2002). En la Figura 8, se presentan los mapas de calor correspondientes a diversos programas de formación, tales como Desarrollo de Software, Animación Digital, Producción Multimedia, Videojuegos, entre otros. Estas visualizaciones permiten establecer la varianza relativa de las variables en cada grupo latente, y aportan una visión comprensiva del comportamiento de los aprendices según el programa. Los resultados evidencian patrones consistentes en múltiples clústeres: por ejemplo, variables como cantidad de quejas y cantidad de reversiones aparecen significativamente elevadas en grupos con mayor riesgo de deserción, especialmente en programas como

Desarrollo de Software, Animación Digital, Gestión de Redes de datos y preprensa. Por otro lado, se identificaron clústeres con valores negativos en dichas variables en programas como Producción de Medios Audiovisuales o Impresión Digital, sugiriendo perfiles más estables. Además, variables demográficas como género y estrato socioeconómico también mostraron varianza significativa en programas como Elaboración de Audiovisuales y Desarrollo de Medios Gráficos Visuales, lo cual plantea la necesidad de implementar estrategias de retención específicas por tipología de programa (Gaitas et al., 2024; Kamata et al., 2018).

Figura 8: Heatmap del v-test para determinar las variables explicativas de los clústeres de cada programa de formación.



Asimismo, se propone la aplicación del Análisis de Correspondencias Múltiples (ACM) a las variables categóricas de cada conjunto de datos por programa de formación, con el objetivo de contrastar y complementar la identificación de las características más relevantes asociadas a los diferentes niveles de deserción. En la Figura 9 se presentan las visualizaciones resultantes para los programas de Análisis y Desarrollo de Software (ADSO) y Videojuegos, donde se representan simultáneamente las variables categóricas y el nivel de deserción. Cabe destacar que este procedimiento se aplicó a la totalidad de los programas incluidos en el estudio.

El contraste de estos resultados con los obtenidos mediante el Análisis de Componentes Principales (PCA) y la Detección de Clases Latentes (DECLA) evidencia la existencia de patrones coincidentes en las características de los clústeres, lo que otorga mayor robustez a las conclusiones.

En ese sentido, en la Tabla 8 se presenta una síntesis de este análisis para cada uno de los programas

evaluados. Dicha tabla permite visualizar los niveles de probabilidad de deserción identificados y propuestos, junto con las características que presentan una mayor relevancia dentro de cada nivel.

Figura 9: Análisis de correspondencias múltiples (ACM) con las variables categóricas de cada uno de los programas de formación analizados.

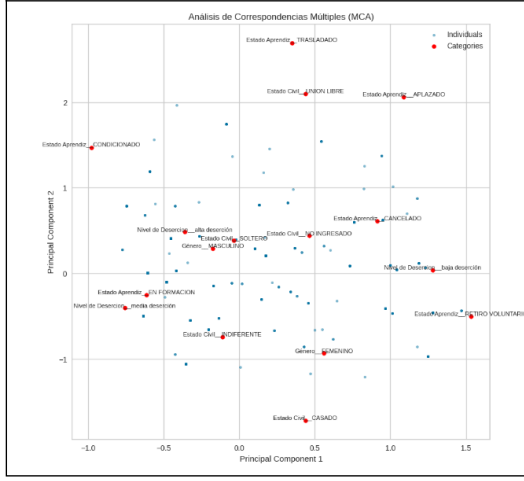
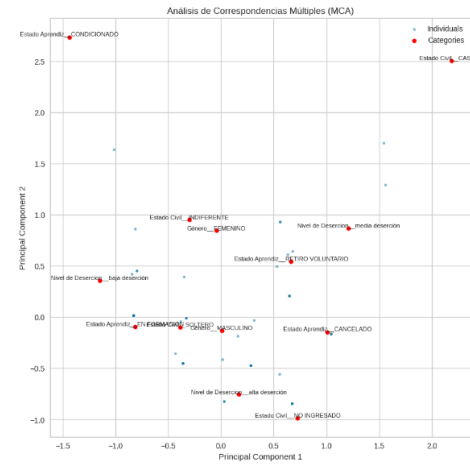
Desarrollo de Software	Videojuegos
	

Tabla 8: Segmentación de Clústeres a partir de la correlación de las características en el PCA.

Programa	Nivel Probable de Deserción		
Desarrollo de Software	**Alta Deserción_C1**	**Media Deserción_C0**	**Baja deserción_C2**
	Altas quejas	Altas Reversiones	Mínimas reversiones
	Genero		Mínimas quejas
	Estado Aprendiz		Estado Aprendiz
Animación 3D	**Alta Deserción_C0**	**Media Deserción_C1**	**Baja deserción_C2**
	Estado Civil	Genero	Mínimas reversiones
	Altas quejas	Edad	Mínimas quejas
	Altas reversiones	Estado Aprendiz	Ubicación
Animación Digital	**Alta Deserción_C2**	**Media Deserción_C1**	**Baja deserción_C0**
	Ubicación	Estado civil	Mínimas reversiones
	Altas quejas	Novedad	Mínimas quejas
	Altas reversiones	Edad	
	Estrato		
Desarrollo de Medios Gráficos Visuales	**Alta Deserción_C1**	**Media Deserción_C0**	**Baja deserción_C2**
	Altas quejas	Estado Civil	Mínimas reversiones
	Altas reversiones	Genero	Mínimas quejas
Elaboración de Audiovisuales	**Alta Deserción_C1**	**Media Deserción_C0**	**Baja deserción_C2**
	Altas quejas	Ubicación	Mínimas reversiones
	Altas reversiones	Estado aprendiz	Mínimas quejas
	Estrato		Edad
			Estado Civil
Gestión de Redes de Datos	**Alta Deserción_C0**	**Media Deserción_C2**	**Baja deserción_C1**
	Altas quejas	Ubicación	Mínimas reversiones
	Altas reversiones		Mínimas quejas
	Genero		
	Estrato		
	Estado Aprendiz		
	Genero		
Impresión Digital	**Alta Deserción_C2**	**Media Deserción_C1**	**Baja deserción_C0**
	Altas quejas	Ubicación	Mínimas reversiones
	Altas reversiones	Estrato	Mínimas quejas
	Novedad	Estado Civil	Genero
			Edad

Tabla 8: Segmentación de Clústeres a partir de la correlación de las características en el PCA.			
Programa	Nivel Probable de Deserción		
Mantenimiento de Equipos de Cómputo	**Alta Deserción_C0**	**Media Deserción_C2**	**Baja deserción_C1**
	Ubicación	Altas Reversiones	Estrato
	Genero	Altas quejas	Estado Civil
	Estado Aprendiz	Género	
Medios Audiovisuales	**Alta Deserción_C2**	**Media Deserción_C1**	**Baja deserción_C0**
	Altas quejas	Altas Reversiones	Altas Reversiones
	Estado Aprendiz	Género	
	Estado Civil	Edad	
		Estrato	
Preprensa	**Alta Deserción_C1**	**Media Deserción_C2**	**Baja deserción_C0**
	Altas Quejas	Estado Civil	Mínimas Quejas
	Altas Reversiones	Estrato	Mínimas Reversiones
	Ubicación	Estado Aprendiz	Edad
	Genero		
Producción Multimedia	**Alta Deserción_C0**	**Media Deserción_C2**	**Baja deserción_C1**
	Altas Quejas	Estado Aprendiz	Mínimas Quejas
	Altas Reversiones	Ubicación	Mínimas Reversiones
			Género
			Edad
			Estrato
Sistemas	**Alta Deserción_C1**	**Media Deserción_C2**	**Baja deserción_C0**
	Altas Reversiones	Estado Civil	Mínimas Reversiones
	Altas Quejas	Genero	Mínimas Quejas
		Estrato	Ubicación
		Estado Aprendiz	Edad
Videojuegos	**Alta Deserción_C2**	**Media Deserción_C0**	**Baja deserción_C1**
	Altas Reversiones	Genero	Mínimas Reversiones
	Altas Quejas	Estrato	Mínimas Quejas
	Ubicación	Edad	Estado Civil
	Estado Aprendiz		

Fuente: Autor

De otro lado, con el fin de determinar el riesgo de deserción estudiantil mediante una estrategia híbrida que combine algoritmos no supervisados y supervisados, como se muestra en la Figura 3, se añade como variable adicional el número de cluster obtenido en la fase de agrupamiento. A partir de esta integración, se construye un dataframe completo con todas las variables originales y la nueva característica de cluster, sobre el cual se efectúa la selección de modelos.

La Tabla 9 presenta los algoritmos con mejor desempeño en este escenario.La incorporación del número de *cluster* como característica representa una técnica avanzada de *feature engineering*, ya que permite capturar patrones latentes que podrían no ser evidentes con las variables iniciales. Estudios recientes han demostrado que este enfoque mejora la calidad de predicción: en entornos educativos, se ha observado que incluir variables derivadas de clustering reduce la necesidad de datos etiquetados y potencia las capacidades predictivas para la detección temprana de deserción (Melchor et al., 2025). Además, investigaciones como (Saad et al., 2025) desarrollan métodos híbridos de detección y *feature engineering* supervisado (SOD FE) que alcanzan F1 > 0.98 en datasets reales, haciendo uso de componentes de agrupamiento y selección de características. Asimismo, Hoca y Dimililer (2025) proponen un marco que primero clasifica alumnos en riesgo y luego los agrupa, facilitando la aplicación de políticas específicas según cada cluster.

Tabla 9: Algoritmos con mejor desempeño con Dataframe Completo.	
Algoritmos	Frecuencia
GaussianNB	5
LogisticRegression	2
GradientBoostingClassifier	2
ExtraTreesClassifier	1
XGBClassifier	1
RandomForestClassifier	1

Fuente: Autor

Asimismo, se estructura un dataframe priorizando, desde una perspectiva orientada al entendimiento del negocio, las características más relevantes para la implementación de una herramienta digital o aplicativo. El objetivo es facilitar la experiencia del usuario, minimizando la cantidad de campos a diligenciar para realizar la predicción. En este sentido, se seleccionan las siguientes variables del dataframe: "Cantidad de quejas", "Cantidad de reversiones", "Género", "Edad", "Estado civil", "Estrato" y "Estado Aprendiz", con el fin de predecir la variable objetivo "Clusters". Posteriormente, se exploran diversos algoritmos de clasificación aplicados a los subconjuntos de datos correspondientes a cada programa de formación. Los modelos que demostraron el mejor desempeño en estos escenarios se presentan en la Tabla 10.

Tabla 10: Algoritmos con mejor desempeño con Dataframe priorizado.

Algoritmos	Frecuencia
RandomForestClassifier	3
GradientBoostingClassifier	3
LinearDiscriminantAnalysis	1
ExtraTreesClassifier	1
LogisticRegression	1
DecisionTreeClassifier	1
GaussianNB	1
LGBMClassifier	1

Fuente: Autor

A partir de los resultados obtenidos en los distintos ejercicios de procesamiento, análisis y selección de modelos aplicados a cada uno de los subconjuntos de datos, se desarrolló una aplicación de predicción como herramienta práctica que integra todo el conocimiento generado durante las fases de preprocesamiento, modelado y validación. Esta aplicación permite predecir tanto el estado del aprendiz como el riesgo de deserción, constituyéndose en un recurso clave para apoyar la toma de decisiones institucionales orientadas a la permanencia estudiantil. La implementación de esta herramienta responde no solo a objetivos analíticos, sino también a criterios de usabilidad y aplicabilidad en contextos reales de gestión educativa. Las interfaces de la aplicación desarrollada se presentan en la Figura 10.

Figura 10: Interfaz de la aplicación desarrollada con apoyo de SCREAMLIT.

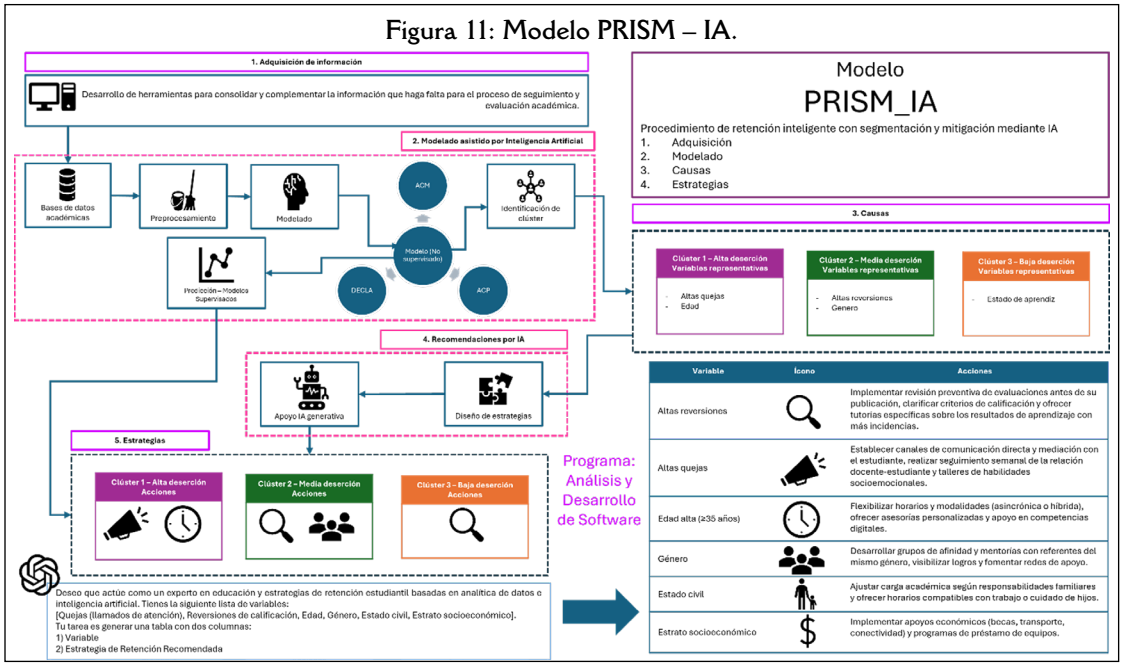
En conjunto, los resultados obtenidos a partir del uso combinado de técnicas de aprendizaje supervisado y no supervisado, junto con el proceso de selección de variables desde una perspectiva analítica y de negocio, evidencian el potencial del enfoque propuesto para la construcción de soluciones predictivas aplicables en contextos educativos. La implementación de la aplicación desarrollada, basada en los modelos con mejor desempeño y en un conjunto optimizado de variables, representará un avance significativo hacia la toma de decisiones fundamentadas en datos, permitiendo no solo anticipar el riesgo de deserción, sino también caracterizar de forma más precisa el estado de los aprendices. Este trabajo no solo valida la pertinencia técnica del enfoque híbrido y del *feature engineering* con *clustering*, sino que también ofrece un recurso concreto para la gestión institucional, contribuyendo activamente a la formulación de estrategias focalizadas de permanencia y acompañamiento estudiantil.

A partir de esta convergencia, se propone un modelo heurístico que vincula características distintivas a cada clúster o nivel de deserción, constituyéndose en un insumo fundamental para el diseño de estrategias de intervención focalizadas y adaptadas a las particularidades de cada programa académico. En este sentido, el presente trabajo permitió estructurar un modelo operativo y formal, que incorpora herramientas de machine learning para desarrollar estrategias de mitigación de la deserción escolar.

Denominado PRISM–IA (Procedimiento de Retención Inteligente con Segmentación y Mitigación mediante Inteligencia Artificial), este procedimiento se plantea como una metodología sistemática para identificar, segmentar y priorizar acciones de intervención, con base en análisis predictivo y caracterización de perfiles de riesgo. Estudios recientes muestran cómo los enfoques de IA segmentan eficazmente a estudiantes en riesgo mediante *clustering* explicable, facilitando políticas de retención adaptadas a cada perfil (Hoca y Dimililer, 2025). Asimismo, se ha evidenciado que combinar datos institucionales con indicadores de compromiso estudiantil (p. ej., interacción en apps, centralidad en redes sociales académicas) aumenta significativamente la precisión predictiva (Matz et al., 2023).

En la Figura 11 se presenta la representación del modelo aplicado a un programa específico; no obstante, su diseño modular y adaptable permite su extensión a otros programas, teniendo en cuenta las propiedades de los clústeres y las predicciones generadas por el aplicativo propuesto.

Figura 11: Modelo PRISM – IA.



La aplicabilidad de este modelo en sistemas curriculares y pedagógicos rígidos radica en su capacidad de operar como un complemento analítico a los procesos tradicionales, sin pretender sustituirlos, sino potenciarlos mediante el monitoreo académico continuo. Para garantizar la precisión y la aplicabilidad

de los modelos, se plantea la integración de indicadores provenientes de los sistemas institucionales de información como registros académicos, tasas de reprobación y patrones de asistencia con métricas de rendimiento generadas por los algoritmos de *Machine Learning*. Esto permitirá un seguimiento dinámico y verificable de las predicciones, reduciendo el margen de error y facilitando la intervención temprana. La necesidad de innovar en este campo se justifica en que, a pesar de los esfuerzos previos y de las medidas preventivas ya implementadas, la persistencia de altas tasas de deserción indica que los enfoques actuales no son suficientes. En este sentido, el modelo propuesto no solo sugiere intervenciones, sino que también plantea la posibilidad de articular nuevas regulaciones institucionales y políticas educativas que respalden el uso sistemático de analítica predictiva en la toma de decisiones. El objetivo central de esta premisa es mostrar que la combinación de enfoques estadísticos y pedagógicos puede abrir un campo de acción más efectivo para la retención, incluso si en esta fase del estudio no se formulan recomendaciones normativas detalladas, dado que el propósito principal es evidenciar la viabilidad técnica y operativa del modelo.

4. Conclusiones

El presente estudio evidenció que la aplicación de algoritmos de aprendizaje automático constituye una estrategia eficaz para la predicción de la deserción estudiantil en instituciones de formación por competencias en Colombia. En particular, el algoritmo Random Forest se destacó por su alta capacidad de precisión y generalización, consolidándose como la técnica con mejor desempeño en la mayoría de los programas analizados.

La combinación de modelos supervisados y no supervisados, a partir de la implementación del Análisis de Componentes Principales (PCA), análisis de correspondencias múltiples (ACM) con las características categóricas, la determinación de clases latentes (DECLA) y la segmentación mediante K-Means, permitió identificar patrones ocultos y categorizar a los estudiantes según niveles de alta, media o baja probabilidad de deserción. Esta estrategia metodológica fortaleció el entendimiento del fenómeno, aportando una caracterización más precisa de los factores asociados al abandono escolar.

La aplicación de la metodología de Detección de Clases Latentes (DECLA), complementada con el análisis estadístico mediante v-test, demostró ser una estrategia robusta para la segmentación e interpretación de datos complejos en entornos educativos. Al analizar múltiples programas de formación del ámbito tecnológico y digital, fue posible identificar patrones latentes de comportamiento estudiantil que no serían evidentes mediante técnicas tradicionales. La caracterización de clústeres permitió evidenciar variables clave asociadas al riesgo de deserción, tales como la cantidad de quejas, número de reinversiones, género y estrato socioeconómico, lo cual ofrece una base empírica para el diseño de intervenciones focalizadas.

Asimismo, la consideración integral de variables sociodemográficas, de seguimiento académico y de desempeño académico, junto con un riguroso proceso de consolidación, preprocesamiento y validación de los datos, mejoró significativamente la capacidad predictiva de los modelos construidos, favoreciendo su aplicabilidad en escenarios educativos reales.

Los hallazgos destacan la pertinencia de incorporar estrategias de intervención temprana basadas en inteligencia artificial como parte de las políticas de retención estudiantil. La articulación de fuentes de datos diversas y técnicas analíticas avanzadas resulta fundamental para diseñar acciones más focalizadas y efectivas que contribuyan a la disminución de los índices de deserción en instituciones de formación por competencias.

Finalmente, el estudio resalta la importancia de diseñar e implementar instrumentos específicos de recolección de datos, como el repositorio de seguimiento académico, el cual resultó ser un insumo clave para el análisis y la clusterización de la población de estudio. La disponibilidad de datos estructurados sobre quejas académicas, reversiones de calificaciones y planes de mejora permitió mejorar la segmentación de los aprendices y fortalecer la precisión de los modelos de predicción, evidenciando la necesidad de contar con sistemas de registro y monitoreo académico más robusto para optimizar los procesos de analítica educativa.

Para fortalecer su efectividad, las instituciones deberían consolidar repositorios estandarizados de datos académicos y sociodemográficos, así como promover la capacitación de directivos y docentes en el uso de analítica educativa como apoyo a la toma de decisiones. De esta forma, los modelos no solo se conciben como ejercicios técnicos, sino como herramientas estratégicas para la permanencia estudiantil en contextos de educación superior.

No obstante, esta investigación presenta ciertas limitaciones. Aunque se incluyeron varios programas de formación, el estudio se circunscribió a una única institución de educación por competencias, lo que

podría limitar la posibilidad de generalizar los resultados a otros contextos. La calidad y disponibilidad de los datos constituyen otro factor determinante, pues la precisión y estabilidad de los modelos puede verse comprometida en escenarios con registros incompletos o inconsistentes. Asimismo, aunque se lograron identificar patrones relevantes, no se evaluó de manera longitudinal la efectividad de las intervenciones derivadas, lo que evidencia la necesidad de futuros estudios que analicen su sostenibilidad en el tiempo. Finalmente, la incorporación de múltiples variables, si bien enriquece la caracterización del fenómeno, también puede incrementar el riesgo de sobreajuste si no se valida con cohortes externas, lo que representa un desafío para investigaciones posteriores orientadas a fortalecer la robustez y transferibilidad del modelo.

Referencias

Al Ka'bi, A. (2023). Proposed artificial intelligence algorithm and deep learning techniques for development of higher education. *International Journal of Intelligent Networks*, *4*, 68-73. https://doi.org/10.1016/j.ijin.2023.03.002

Albán, M. y Mauricio, D. (2018). Decision Trees for the Early Identification of University Students at Risk of Desertion. *International Journal of Engineering and Technology*, *7*(4.44), 51-54. https://doi.org/10.14419/ijet.v7i4.44.26862

Ali, J. A., Muse, A. H., Abdi, M. K., Ali, T. A., Muse, Y. H. y Cumar, M. A. (2025). Machine learning-driven analysis of academic performance determinants: Geographic, socio-demographic, and subject-specific influences in Somaliland's 2022–2023 national primary examinations. *International Journal of Educational Research Open*, *8*, 100426. https://doi.org/10.1016/j.ijedro.2024.100426

Améstica-Rivas, L., King-Domínguez, A., Sanhueza Gutiérrez, D. A. y Ramírez González, V. (2021). Efectos económicos de la deserción en la gestión universitaria: el caso de una universidad pública chilena. *Hallazgos*, *18*(35), 209-231. https://doi.org/10.15332/2422409X.5772

Andrés Rico, P. (2022). Modelos predictivos progresivos del rendimiento académico de estudiantes universitarios. *RIDE. Revista Iberoamericana para la Investigación y el Desarrollo Educativo*, *12*(24), e350. https://doi.org/10.23913/ride.v12i24.1196

Arnaud Bobadilla, A. J., Sánchez Villarreal, F., Galindo Miranda, N. E., Franco Bodek, D. y Ruiz Gutiérrez, R. (2022). Diagnóstico de las causas de rezago y deserción en alumnos de la Facultad de Ciencias de la UNAM. *RIDE. Revista Iberoamericana para la Investigación y el Desarrollo Educativo*, *12*(24), e342. https://doi.org/10.23913/ride.v12i24.1181

Ayuso del Puerto, D. y Gutiérrez Esteban, P. (2022). La Inteligencia Artificial como recurso educativo durante la formación inicial del profesorado. *RIED-Revista Iberoamericana de Educación a Distancia*, *25*(2), 347-362. https://doi.org/10.5944/ried.25.2.32332

Barrios-Tao, H., Díaz, V. y Guerra, Y. M. (2021). Propósitos de la educación frente a desarrollos de inteligencia artificial. *Cadernos de Pesquisa*, *51*, e07767. https://doi.org/10.1590/198053147767

Barrios, I. (2023). Artificial intelligence and scientific writing: ethical aspects in the use of new technologies. *Medicina Clínica y Social*, *7*(2), 46-47. https://doi.org/10.52379/mcs.v7i2.278

Bitencourt, W. A., Silva, D. M. y Xavier, G. d. C. (2022). ¿Puede la inteligencia artificial apoyar acciones contra la deserción escolar universitaria? *Ensaio: Avaliação e Políticas Públicas em Educação*, *30*(116), 669-694. https://doi.org/10.1590/S0104-403620220003002854

Bolaño-García, M. y Duarte-Acosta, N. (2024). Una revisión sistemática del uso de la inteligencia artificial en la educación. *Revista Colombiana de Cirugía*, *39*(1), 51-63. https://doi.org/10.30944/20117582.2365

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

Bressane, A., Spalding, M., Zwirn, D., Loureiro, A. I. S., Bankole, A. O., Negri, R. G., de Brito Junior, I., et al. (2022). Fuzzy Artificial Intelligence—Based Model Proposal to Forecast Student Performance and Retention Risk in Engineering Education: An Alternative for Handling with Small Data. *Sustainability*, *14*(21), 14071. https://doi.org/10.3390/su142114071

Bustamante Bula, R. y Camacho Bonilla, A. (2024). Inteligencia artificial (IA) en las escuelas: una revisión sistemática (2019-2023). *Enunciación*, *29*(1), 62-82. https://doi.org/10.14483/22486798.22039

Cabrales, R. A., Moreno, G. A., Trujillo, S. E., Londoño, J. F. y Patiño, V. M. (2022). Deserción, rezago estudiantil y egreso exitoso en 40 cohortes del Programa de Medicina de la Universidad Tecnológica de Pereira. Colombia. *Iatreia*, *35*(3), 239-248. https://doi.org/10.17533/10.17533/udea.iatreia.133

Caceres-Correa, I. (2021). Acerca de la escolaridad a distancia y la deserción en pandemia. *Utopía y Praxis Latinoamericana*, *26*(2), 11-12. https://www.redalyc.org/journal/279/27966514001/27966514001.pdf

Camargo García, A. J. (2020). *Modelo para la predicción de la deserción de estudiantes de pregrado, basado en técnicas de minería de datos* [Trabajo de grado - Maestría, Corporación Universidad de la Costa]. https://hdl.handle.net/11323/7077

Cardozo, S., Silveira, A. y Fonseca, B. (2022). Detección temprana del riesgo escolar. Predicción de trayectorias de rezago en la educación primaria en Uruguay mediante técnicas de machine learning. *Revista Latinoamericana de Estudios Educativos*, *52*(2), 297-326. https://doi.org/10.48102/rlee.2022.52.2.391

Castillejos López, B. (2022). Inteligencia artificial y los entornos personales de aprendizaje: atentos al uso adecuado de los recursos tecnológicos de los estudiantes universitarios. *Educación*, *31*(60), 9-24. https://doi.org/10.18800/educacion.202201.001

Castrillón, O. D., Sarache, W. y Ruiz-Herrera, S. (2020). Predicción del rendimiento académico por medio de técnicas de inteligencia artificial. *Formación Universitaria*, *13*(1), 93-102. https://doi.org/10.4067/S0718-50062020000100093

Castro-Maldonado, J. J., Patiño-Murillo, J. A. y Camargo-Casallas, E. (2022). Aplicación de analítica de datos en la evaluación de los procesos de investigación aplicada y desarrollo experimental para fortalecer las competencias del siglo XXI en una institución de educación no formal. *Respuestas*, *27*(2), 6-26. https://doi.org/10.22463/0122820X.3541

Chalpartar Nasner, L. T. M., Fernández Guzmán, A. M., Betancourth Zambrano, S. y Gómez Delgado, Y. A. (2022). Deserción en la población estudiantil universitaria durante la pandemia, una mirada cualitativa. *Revista Virtual Universidad Católica del Norte*, (66), 37-62. https://doi.org/10.35575/rvucn.n66a3

Christou, V., Tsoulos, I., Loupas, V., Tzallas, A. T., Gogos, C., Karvelis, P. S., Antoniadis, N., et al. (2023). Performance and early drop prediction for higher education students using machine learning. *Expert Systems with Applications*, *225*, 120079. https://doi.org/10.1016/j.eswa.2023.120079

Contreras Bravo, L. E., Nieves-Pimiento, N. y Gonzalez-Guerrero, K. (2022). Prediction of University-Level Academic Performance through Machine Learning Mechanisms and Supervised Methods. *Ingeniería*, *28*(1), e19514. https://doi.org/10.14483/23448393.19514

Coto Jiménez, M. (2021). Consideraciones para la incorporación de la Inteligencia Artificial en un programa de pregrado de Ingeniería Eléctrica. *Actualidades Investigativas en Educación*, *21*(2), 529-555. https://doi.org/10.15517/aie.v21i2.44893

de Freitas e Silva, P. T. y Bezerra Sampaio, L. M. (2022). Student retention policies in higher education: reflections from a literature review for the Brazilian context. *Revista de Administração Pública*, *56*(5), 603-631. https://doi.org/10.1590/0034-761220220034x

Delogu, M., Lagravinese, R., Paolini, D. y Resce, G. (2024). Predicting dropout from higher education: Evidence from Italy. *Economic Modelling*, *130*, 106583. https://doi.org/10.1016/j.econmod.2023.106583

Di Paola Naranjo, A., Sánchez, S. y Pereno, G. L. (2022). Factores sociodemográficos que inciden en la retención de ingresantes a la universidad: un estudio exploratorio en la Licenciatura en Psicología de la Universidad Nacional de Córdoba (UNC). *Revista Educación*, *46*(2), 209-226. https://doi.org/10.15517/revedu.v46i2.47784

Fernández-Martín, T., Solís-Salazar, M., Hernández-Jiménez, M. T. y Moreira-Mora, T. E. (2019). A Multinomial and Predictive Analysis of Factors Associated with University Dropout. *Revista Electrónica Educare*, *23*(1), 73-97. https://doi.org/10.15359/ree.23-1.5

Flores, V., Heras, S. y Julian, V. (2022). Comparison of Predictive Models with Balanced Classes Using the SMOTE Method for the Forecast of Student Dropout in Higher Education. *Electronics*, *11*(3), 457. https://doi.org/10.3390/electronics11030457

Forero-Corba, W. y Negre Bennasar, F. (2024). Técnicas y aplicaciones del Machine Learning e inteligencia artificial en educación: una revisión sistemática. *RIED-Revista Iberoamericana de Educación a Distancia*, *27*(1), 209-253. https://doi.org/10.5944/ried.27.1.37491

Fuertes Arroyo, Y. N. y Uc Ríos, C. E. (2023). Aporte de las tecnologías de la información y la comunicación (TIC) para minimizar la deserción de carreras universitarias en tecnología. *Revista Virtual Universidad Católica del Norte*, (68), 4-36. https://doi.org/10.35575/rvucn.n68a2

Gaitas, S., Silva, J. C. y Poças, A. (2024). A latent class analysis on students' beliefs about teachers' practices enhancing their well-being. *Frontiers in Education*, *9*, 1252222. https://doi.org/10.3389/feduc.2024.1252222

García Esquirol, Ó. (2015). Futuro de la enseñanza médica: inteligencia artificial y big data. *FEM: Revista de la Fundación Educación Médica*, *18*, s60-s61. https://doi.org/10.4321/S2014-98322015000300009

García Peñalvo, F. J., Llorens-Largo, F. y Vidal, J. (2024). La nueva realidad de la educación ante los avances de la inteligencia artificial generativa. *RIED-Revista Iberoamericana de Educación a Distancia*, *27*(1), 9-39. https://doi.org/10.5944/ried.27.1.37716

Gil-Vera, V. D. y Quintero-López, C. (2021). Predicción del rendimiento académico estudiantil con redes neuronales artificiales. *Información tecnológica*, *32*(6), 221-228. https://doi.org/10.4067/S0718-07642021000600221

Gonzalez Salas Duhne, P., Delgadillo, J. y Lutz, W. (2022). Predicting early dropout in online versus face-to-face guided self-help: A machine learning approach. *Behaviour Research and Therapy*, *159*, 104200. https://doi.org/10.1016/j.brat.2022.104200

Gual-Sala, A. (2023). La inteligencia artificial y la educación médica (I): la revolución profesional. *FEM: Revista de la Fundación Educación Médica*, *26*(2), 43-47. https://doi.org/10.33588/fem.262.1271

Gutierrez-Pachas, D. A., Garcia-Zanabria, G., Cuadros-Vargas, E., Camara-Chavez, G. y Gomez-Nieto, E. (2023). Supporting Decision-Making Process on Higher Education Dropout by Analyzing Academic, Socioeconomic, and Equity Factors through Machine Learning and Survival Analysis Methods in the Latin American Context. *Education Sciences*, *13*(2), 154. https://doi.org/10.3390/educsci13020154

Guzmán-Castillo, S., Körner, F., Pantoja-García, J. I., Nieto-Ramos, L., Gómez-Charris, Y., Castro-Sarmiento, A. y Romero-Conrado, A. R. (2022). Implementation of a Predictive Information System for University Dropout Prevention. *Procedia Computer Science*, *198*, 566-571. https://doi.org/10.1016/j.procs.2021.12.287

Henriquez Cabezas, N. y Vargas Escobar, D. (2022). Modelos predictivos de rendimiento y deserción académica en estudiantes de primer año de una universidad pública chilena. *Revista de Estudios y Experiencias en Educación*, *21*(45), 299-316. https://doi.org/10.21703/0718-5162.v21.n45.2022.015

Hernández-Medina, P. y Ramírez-Torres, G. (2022). Evaluación de impacto del financiamiento educativo en la deserción y la graduación: un análisis de regresiones discontinuas. *Revista Iberoamericana de Educación Superior*, *13*(37), 63-82. https://doi.org/10.22201/iisue.20072872e.2022.37.1304

Hernández Arias, A. (2023). La Inteligencia Artificial como herramienta de apoyo en las actividades de investigación. *Compendium*, *26*(50), 1. https://doi.org/10.5281/zenodo.10268867

Hidalgo Suarez, C. G., Bucheli-Guerrero, V. A. y Ordóñez-Eraso, H. A. (2023). Artificial Intelligence and Computer-Supported Collaborative Learning in Programming: A Systematic Mapping Study. *Tecnura*, *27*(75), 175-206. https://doi.org/10.14483/22487638.19637

Hoca, S. y Dimililer, N. (2025). A Machine Learning Framework for Student Retention Policy Development: A Case Study. *Applied Sciences*, *15*(6), 2989. https://doi.org/10.3390/app15062989

Hoyos Osorio, J. K. y Daza Santacoloma, G. (2023). Predictive model to identify college students with high dropout rates. *Revista Electrónica de Investigación Educativa*, *25*, e13. https://doi.org/10.24320/redie.2023.25.e13.5398

Ibarra-Vazquez, G., Ramírez-Montoya, M. S., Buenestado-Fernández, M. y Olague, G. (2023). Predicting open education competency level: A machine learning approach. *Heliyon*, *9*(11), e20597. https://doi.org/10.1016/j.heliyon.2023.e20597

Incio-Flores, F. A., Capuñay-Sanchez, D. L. y Estela-Urbina, R. O. (2023). Artificial Neural Network Model to Predict Academic Results in Mathematics II. *Revista Electrónica Educare*, *27*(1), 1-19. https://doi.org/10.15359/ree.27-1.14516

Jalón Arias, E., Ponce Ruiz, D., Arandia, J. C. y Arrias Añez, J. C. (2021). Las limitaciones de la aplicación de la inteligencia artificial al derecho y el futuro de la educación jurídica. *Conrado*, *17*(83), 439-450. https://conrado.ucf.edu.cu/index.php/conrado/article/view/2116

Jimenez Chaves, V. E. y García Torres, M. (2019). Análisis de la Educación Inicial en Paraguay a través de las Técnicas de Aprendizaje Automático. *Revista de la Sociedad Científica del Paraguay*, *24*(2), 293-304. https://doi.org/10.32480/rscp.2019-24-2.293-304

Juca-Maldonado, F. (2023). El impacto de la inteligencia artificial en los trabajos académicos y de investigación. *Revista metropolitana de Ciencias aplicadas*, *6*(Suplemento 1), 289-296. https://doi.org/10.62452/8nww1k83

Kamata, A., Kara, Y., Patarapichayatham, C. y Lan, P. (2018). Evaluation of Analysis Approaches for Latent Class Analysis with Auxiliary Linear Growth Model. *Frontiers in Psychology*, *9,* 130. https://doi.org/10.3389/fpsyg.2018.00130

Kordbagheri, A., Kordbagheri, M., Tayim, N., Fakhrou, A. y Davoudi, M. (2025). Using advanced machine learning algorithms to predict academic major completion: A cross-sectional study. *Computers in Biology and Medicine*, *184,* 109372. https://doi.org/10.1016/j.compbiomed.2024.109372

Krüger, J. G. C., Britto, A. d. S. y Barddal, J. P. (2023). An explainable machine learning approach for student dropout prediction. *Expert Systems with Applications*, *233*, 120933. https://doi.org/10.1016/j.eswa.2023.120933

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

Londoño-Gallego, J. A., Andrade-Martelo, I. C., Castro-Maldonado, J. J. y Reyes-Moreno, E. R. (2024). Aplicación de ChatGPT como innovación educativa en los procesos de enseñanza y aprendizaje en la formación por competencias: Un análisis aplicando técnicas de Machine Learning. *Respuestas*, *29*(1), 52-66. https://doi.org/10.22463/0122820X.4254

Lopezosa, C. (2023). Generative Artificial Intelligence in Scientific Communication: Challenges and Opportunities. *Revista de Investigacion E Innovacion en Ciencias de la Salud*, *5*(1), 1-5. https://doi.org/10.46634/riics.211

Lozano Treviño, D. y Maldonado Maldonado, L. (2022). Asociación entre factores institucionales y escolares con la propensión de deserción escolar en colegios militarizados. *Revista de Estudios y Experiencias en Educación*, *21*(47), 287-306. https://doi.org/10.21703/0718-5162202202102147016

Magidson, J. y Vermunt, J. (2002). Latent class models for clustering: A comparison with K-means. *Canadian Journal of Marketing Research*, *20*(1), 36-43. https://jeroenvermunt.nl/cjmr2002.pdf

Marrón Ramos, D. N., Reyes Valenzuela, R., González Torres, A., Juárez Rodríguez, R. y Mendoza Montero, F. Y. (2022). Evaluación de la deserción a nivel superior: dimensiones que inciden en carreras universitarias. *RIDE. Revista Iberoamericana para la Investigación y el Desarrollo Educativo*, *13*(25). https://doi.org/10.23913/ride.v13i25.1269

Martínez, J. y Castillo, D. (2024). Prediction of student dropout using Artificial Intelligence algorithms. *Procedia Computer Science*, *251*, 764-770. https://doi.org/10.1016/j.procs.2024.11.182

Matz, S. C., Bukow, C. S., Peters, H., Deacons, C., Dinu, A. y Stachl, C. (2023). Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics. *Scientific Reports*, *13*(1), 5705. https://doi.org/10.1038/s41598-023-32484-w

Melchor, F., Conejero, J. M., Fernández-García, A. J., Sánchez-Figueroa, F. y Rodríguez-Echeverría, R. (2025). An empirical evaluation of clustering processes for early detection of university dropout. *Research Square*. https://doi.org/10.21203/rs.3.rs-6146415/v1

Múnera-Duque, A. (2023). Inteligencia artificial y cirugía. *Revista Colombiana de Cirugía*, *38*(2), 231-232. https://doi.org/10.30944/20117582.2341

Mustofa, S., Emon, Y. R., Mamun, S. B., Akhy, S. A. y Ahad, M. T. (2025). A novel AI-driven model for student dropout risk analysis with explainable AI insights. *Computers and Education: Artificial Intelligence*, *8,* 100352. https://doi.org/10.1016/j.caeai.2024.100352

Navarro Roldán, C. P. y Zamudio Sisa, L. E. (2021). Cuestionario de riesgo de deserción universitaria (CDUe) basado en el modelo ecológico. *Tesis Psicológica*, *16*(1), 244-263. https://doi.org/10.37511/tesis.v16n1a12

Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E. y Nshimyumukiza, P. C. (2022). Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. *Computers and Education: Artificial Intelligence*, *3,* 100066. https://doi.org/10.1016/j.caeai.2022.100066

Ocaña-Fernández, Y., Valenzuela-Fernández, L. A. y Garro-Aburto, L. L. (2019). Inteligencia artificial y sus implicaciones en la educación superior. *Propósitos y Representaciones*, *7*(2), 536-568. https://doi.org/10.20511/pyr2019.v7n2.274

OECD. (2025, Septiembre 9). *Education at a Glance 2025: Colombia*. https://www.oecd.org/en/publications/education-at-a-glance-2025_1a3543e2-en/colombia_32e069b7-en.html

Parkavi, R., Karthikeyan, P. y Abdullah, A. S. (2023). Predicting academic performance of learners with the three domains of learning data using neuro-fuzzy model and machine learning algorithms. *Journal of Engineering Research*, *12*(3), 397–411. https://doi.org/10.1016/j.jer.2023.09.006

Pereira Santana, A. E. y Vidal Cortez, M. (2020). Deserción estudiantil en la educación superior: reflexiones sobre la gestión enfocada en la retención o la permanencia. *Revista Educación*, *45*(1), 519-533. https://doi.org/10.15517/revedu.v45i1.40602

Pineda-Pertuz, C. M., Martinez, Y. y Diaz, I. (2022). Machine Learning algorithms to predict desertion in the faculty of Engineering Sciences at the Corporación Universitaria Antonio José de Sucre. *IOP Conference Series: Materials Science and Engineering*, *1253*(1), 012013. https://doi.org/10.1088/1757-899X/1253/1/012013

Rabelo, A. M. y Zárate, L. E. (2025). A model for predicting dropout of higher education students. *Data Science and Management*, *8*(1), 72-85. https://doi.org/10.1016/j.dsm.2024.07.001

Rodríguez Almazán, Y., Parra-González, E. F., Zurita-Aguilar, K. A., Mejía Miranda, J. y Bonilla Carranza, D. (2023). ChatGPT: La inteligencia artificial como herramienta de apoyo al desarrollo de las competencias STEM en los procesos de aprendizaje de los estudiantes. *ReCIBE. Revista electrónica de Computación, Informática, Biomédica y Electrónica*, *12*(1), C5-12. https://doi.org/10.32870/recibe.v12i1.291

Rodríguez Chávez, M. H. (2021). Sistemas de tutoría inteligente y su aplicación en la educación superior. *RIDE. Revista Iberoamericana para la Investigación y el Desarrollo Educativo*, *11*(22), e175. https://doi.org/10.23913/ride.v11i22.848

Rodríguez, P., Villanueva, A., Dombrovskaia, L. y Valenzuela, J. P. (2023). A methodology to design, develop, and evaluate machine learning models for predicting dropout in school systems: the case of Chile. *Education and Information Technologies*, *28*(8), 10103-10149. https://doi.org/10.1007/s10639-022-11515-5

Saad, S., Yakut, Ö., Alzabidi, E. y Fındık, O. (2025). SOD-FE: A Supervised Outlier Detection and Feature Engineering Approach for Student Dropout Prediction. *Research Square*. https://doi.org/10.21203/rs.3.rs-6889300/v1

Sanhueza Gutiérrez, D., King-Domínguez, A. y Améstica-Rivas, L. (2021). Incidencia de la gestión universitaria en la deserción estudiantil de las universidades públicas en Chile. *IE Revista de Investigación Educativa de la REDIECH*, *12*, e1270. https://doi.org/10.33010/ie_rie_rediech.v12i0.1270

Smith Uldall, J. y Gutiérrez Rojas, C. (2022). An Application of Machine Learning in Public Policy: Early Warning Prediction of School Dropout in the Chilean Public Education System. *Multidisciplinary Business Review*, *15*(1), 20-35. https://doi.org/10.35692/07183992.15.1.4

Tete, M. F., Sousa, M. d. M., de Santana, T. S. y Silva, S. F. (2022). Aplicação de métodos preditivos em evasão no ensino superior: Uma revisão sistemática da literatura. *Education Policy Analysis Archives*, *30*(149). https://doi.org/10.14507/epaa.30.6845

Treviño, M., Ibarra, S., Castán, J., Laria, J. y Guzmán, J. (2013). A Framework to avoid Scholar Desertion using Artificial Intelligence. En *Proceedings of the World Congress on Engineering* (Vol. 3, pp. 1493-1497). International Association of Engineers London. https://www.iaeng.org/publication/WCE2013/WCE2013_pp1493-1497.pdf

Valero Cajahuanca, J. E., Navarro Raymundo, Á. F., Larios Franco, A. C. y Julca Flores, J. D. (2022). Deserción universitaria: Evaluación de diferentes algoritmos de Machine Learning para su predicción. *Revista de Ciencias Sociales*, *28*(3), 362-375. https://doi.org/10.31876/rcs.v28i3.38480

Vidal Ledo, M. J., Madruga González, A. y Valdés Santiago, D. (2019). Inteligencia artificial en la docencia médica. *Educación Médica Superior*, *33*(3). http://scielo.sld.cu/pdf/ems/v33n3/1561-2902-ems-33-03-e1970.pdf

Viloria, A., Pineda Lezama, O. B. y Varela, N. (2019). Bayesian Classifier Applied to Higher Education Dropout. *Procedia Computer Science*, *160*, 573-577. https://doi.org/10.1016/j.procs.2019.11.045

Zimányi, K., Montes Ortiza, A., Houde, P. M. A. y Richter, K. G. (2022). Deserción escolar en dos licenciaturas en la enseñanza de lenguas: un estudio de caso en una universidad pública en México. *Revista de la Educación Superior*, *51*(203), 89-116. https://doi.org/10.36857/resu.2022.203.2220