Modelos predictivos en la orientación vocacional: un análisis teórico para la reducción de la deserción universitaria

Carlo Mario Avila Muñoz

Asesor

Sixyel Jeyson Castaneda Coronado

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Sociales Artes y Humanidades ECSAH

Especialización en Ciencia de Datos y Analítica

2026

Dedicatoria

A mis padres, Mario Avila y Pilar Muñoz, por su gran amor y apoyo incondicional.

A mi hermano, Isaac Avila, por estar siempre dispuesto a apoyarme con todos los impedimentos que surgieron.

Agradecimientos

Quiero agradecer primeramente a Dios que me da la salud y la disposición para levantarme cada día. También quiero expresar mi más sincero agradecimiento a la Universidad Nacional Abierta y a Distancia, por brindarme las herramientas y los contenidos necesarios para crecer académicamente. Al profesor Syxyel, director de este trabajo, gracias por sus consejos oportunos y por enseñarme a como enfocarme.

Resumen

Esta monografía de tipo compilatorio, con enfoque descriptivo-analítico, tiene como objetivo analizar los modelos predictivos basados en técnicas de aprendizaje supervisado aplicados al ámbito de la orientación vocacional, con el fin de contribuir a la reducción de la deserción universitaria. A través de una revisión sistemática de literatura científica en bases de datos académicas como Scopus, IEEE, Google Scholar, UNAD, se identificaron y clasificaron 35 estudios que emplean técnicas de minería de datos como regresión logística, árboles de decisión, Random Forest, XGBoost y redes neuronales. Se examinan las variables académicas, socioeconómicas y vocacionales, así como las métricas de evaluación (exactitud, sensibilidad, precisión, F1-score, AUC-ROC) que permiten validar su desempeño. Los resultados muestran que Random Forest y la regresión logística son las técnicas más utilizadas, con un rendimiento que oscila entre el 70% y el 85% de precisión, según el contexto. Se identifica como principal brecha la escasa incorporación de variables vocacionales (solo en el 23% de los estudios). Finalmente, se proponen lineamientos teóricos para la implementación ética y efectiva de estos modelos como herramienta de apoyo a la orientación vocacional, destacando la necesidad de equilibrar precisión e interpretabilidad, y de integrar constructos psicoeducativos como autoeficacia, indecisión y adaptabilidad profesional.

***Palabras clave*:** Orientación vocacional; análisis de datos; deserción estudiantil; modelos predictivos; aprendizaje supervisado; minería de datos educativa

Abstract

This compilatory monograph, with a descriptive-analytical approach, aims to analyze predictive models based on supervised learning techniques applied to the field of vocational orientation, in order to contribute to the reduction of university dropout. Through a systematic review of scientific literature in academic databases (Scopus, IEEE, Google Scholar, UNAD), 35 studies employing data mining techniques such as logistic regression, decision trees, Random Forest, XGBoost, and neural networks were identified and classified. The variables academic, socioeconomic, and vocational are examined, as well as the evaluation metrics (accuracy, recall, precision, F1-score, AUC-ROC) used to validate their performance. The results show that Random Forest and logistic regression are the most frequently used techniques, with performance ranging between 70% and 85% accuracy, depending on the context. A major gap identified is the limited inclusion of vocational variables (only 23% of the studies). Finally, theoretical guidelines are proposed for the ethical and effective implementation of these models as a support tool for vocational orientation, highlighting the need to balance accuracy and interpretability, and to integrate psychoeducational constructs such as self-efficacy, indecision, and career adaptability.

***Keywords*:** vocational orientation; data analysis; student dropout; predictive models; supervised learning; educational data mining

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Lista de Apéndices

Introducción

La transición de la educación media a la educación superior representa uno de los momentos más críticos en la vida de los jóvenes, no solo por las implicaciones personales y familiares que conlleva, sino también por su impacto en el desarrollo social y económico de las naciones. En este proceso, la elección de una carrera universitaria se erige como una decisión de alta complejidad, cargada de incertidumbre y condicionada por múltiples factores tales como intereses personales, presiones familiares, expectativas salariales, percepciones sociales sobre las profesiones y el propio desconocimiento de las aptitudes individuales (Rodríguez-Muñiz et al., 2019). Cuando esta decisión no se fundamenta en un adecuado proceso de autoconocimiento y de información sobre las opciones disponibles, aumenta significativamente el riesgo de insatisfacción académica y, en última instancia, de deserción universitaria. (Navarro Guzmán & Casero Martinez, 2012)

La deserción en la educación superior es un fenómeno de carácter multicausal que afecta tanto a instituciones públicas como privadas, generando pérdidas económicas significativas para las familias, el Estado y las propias universidades, así como profundas consecuencias emocionales y profesionales para los estudiantes que abandonan sus estudios (Facundo Diaz, 2009) En el contexto colombiano, y particularmente en modalidades de educación a distancia como la ofrecida por la Universidad Nacional Abierta y a Distancia UNAD, la deserción se ha convertido en un desafío recurrente que ha motivado múltiples estudios orientados a comprender sus causas y a proponer estrategias de mitigación (Facundo Diaz, 2009

En respuesta a esta problemática, según (Avila Pérez, 2021) los avances tecnológicos y el auge de la ciencia de datos han abierto nuevas posibilidades para apoyar los procesos de orientación vocacional. En los últimos años, diversas investigaciones han explorado la aplicación

de técnicas de minería de datos y aprendizaje automático con el propósito de predecir la deserción estudiantil, identificar estudiantes en riesgo y, de manera más incipiente, recomendar carreras universitarias con base en perfiles de intereses y aptitudes (Heredia et al., 2015). Los modelos predictivos, particularmente aquellos basados en aprendizaje supervisado, han demostrado ser herramientas prometedoras para procesar grandes volúmenes de datos y generar patrones que apoyen la toma de decisiones en contextos educativos. (Avila Pérez & Medina, 2020)

En este contexto, el presente trabajo de grado se desarrolla como una monografía de revisión con enfoque descriptivo-analítico, cuyo propósito central es analizar los modelos predictivos basados en técnicas de aprendizaje supervisado aplicados al ámbito de la orientación vocacional. A través de una revisión sistemática de la literatura científica en bases de datos académicas, se busca identificar, clasificar y comparar las principales técnicas empleadas, las variables consideradas, las métricas de evaluación utilizadas y los resultados reportados en estudios previos.

La estructura del documento se organiza en cinco capítulos principales. En el primer capítulo, se aborda el planteamiento del problema, presentando la pregunta de investigación que orienta el estudio. El segundo capítulo desarrolla el marco conceptual y teórico, abordando conceptos fundamentales como la deserción universitaria, la orientación vocacional, la minería de datos y el aprendizaje supervisado, así como las principales métricas de evaluación de modelos predictivos. El tercer capítulo expone la metodología empleada, detallando las fases de la revisión sistemática, los criterios de selección de fuentes y las estrategias de análisis. El cuarto capítulo presenta los resultados obtenidos, organizados en torno a la clasificación de modelos, el análisis de variables y la comparación de métricas de desempeño. Finalmente, el quinto capítulo

ofrece una discusión crítica de los hallazgos, las conclusiones del estudio y una propuesta de lineamientos teóricos para la implementación de modelos predictivos como herramienta de apoyo en la orientación vocacional, contribuyendo así a la reflexión académica sobre cómo la analítica de datos puede ayudar a reducir la incertidumbre y mitigar la deserción universitaria.

Planteamiento del Problema

La elección de una carrera universitaria constituye una de las decisiones más trascendentales en la trayectoria de vida de los jóvenes, no solo por su impacto en la realización personal y profesional, sino también por las implicaciones económicas y sociales que conlleva. (Rodriguez Ramirez et al., 2022) señalan que este proceso de elección está influenciado por múltiples factores tales como intereses personales, presiones familiares, expectativas salariales y percepciones sociales sobre las profesiones y, cuando no se fundamenta en un adecuado autoconocimiento, aumenta significativamente el riesgo de insatisfacción académica y deserción. En la misma línea, (Navarro Guzmán & Casero Martinez, 2012) evidencian que las diferencias de género y la falta de información sobre las opciones disponibles contribuyen a decisiones vocacionales poco acertadas.

El fenómeno de la deserción universitaria, entendido como el abandono definitivo de los estudios antes de su culminación (Tinto, 1993), afecta tanto a instituciones públicas como privadas, generando pérdidas económicas para las familias, el Estado y las universidades, así como profundas consecuencias emocionales para los estudiantes que abandonan (Facundo Diaz, 2009). En el contexto colombiano, y particularmente en la modalidad de educación a distancia ofrecida por la UNAD, las tasas de deserción han sido históricamente elevadas, motivando estudios orientados a comprender sus causas y a proponer estrategias de mitigación (Abadía et al., 2018;).

Frente a esta problemática, los avances en minería de datos y aprendizaje automático han abierto nuevas posibilidades para apoyar los procesos de orientación vocacional. Diversos autores han explorado la aplicación de técnicas predictivas para identificar estudiantes en riesgo de deserción (Dake & Buabeng-Andoh, 2022; Heredia et al., 2015; Kemper et al., 2020) y, de

manera más incipiente, para recomendar carreras con base en perfiles de intereses y aptitudes (Mahboob et al., 2024; Rodriguez Ramirez et al., 2022). Sin embargo, la mayoría de estos estudios se centran en variables académicas y socioeconómicas, dejando de lado constructos centrales de la orientación vocacional como la autoeficacia, la indecisión profesional y la adaptabilidad de carrera (Lent et al., 1994; Savickas, 2019; Zhu et al., 2019).

Esta desconexión entre el desarrollo técnico de los modelos predictivos y su integración con los fundamentos psicoeducativos de la orientación vocacional limita su potencial para abordar las causas profundas del desajuste carrera-estudiante. Por ello, la presente investigación se pregunta: *¿Cuáles son las principales técnicas de modelado predictivo basadas en aprendizaje supervisado que pueden aplicarse en el ámbito de la orientación vocacional, y cuáles son sus ventajas, limitaciones y lineamientos de implementación para contribuir a la reducción de la deserción universitaria?*

Justificación

La elección de una carrera universitaria es una decisión crucial que impacta significativamente el desarrollo personal y profesional de los individuos, así como el progreso social y económico de un país. La falta de información adecuada, la presión social, el desconocimiento de las propias aptitudes y la incertidumbre sobre el futuro son factores que con frecuencia conducen a una elección poco acertada. Esta situación, a menudo, deriva en el abandono prematuro de los estudios superiores, generando pérdidas de tiempo, recursos económicos y afectando la motivación de los estudiantes. (García-Botero et al., 2022)

Si bien los avances tecnológicos han abierto nuevas posibilidades para abordar esta problemática a través de sistemas de apoyo a la decisión, existe una brecha entre el desarrollo técnico de estos modelos y su implementación efectiva en contextos educativos reales. Por lo tanto, surge la necesidad de comprender, desde una perspectiva teórica y analítica, cuáles son las técnicas de modelado predictivo más pertinentes, cuáles son sus fortalezas y limitaciones, y cómo pueden integrarse de manera efectiva en los procesos de orientación vocacional.

Investigaciones recientes han ampliado el espectro de aplicación de los modelos predictivos en contextos educativos diversos. Por ejemplo, (O'neill, 2024) empleó Random Forest para predecir la retención de estudiantes del programa Educational Opportunity Fund, mientras que (Orozco-Rodríguez et al., 2025) utilizaron regresión logística para analizar la deserción en programas de ingeniería. En entornos mediados por tecnología, (Rebelo Marcolino et al., 2025) aplicaron CatBoost sobre registros de interacción en Moodle, y (Gochhayat & Ravindran, 2025) recurrieron también a la regresión logística para estudiar la deserción escolar en India. Desde una perspectiva más amplia, la revisión sistemática de (García-Botero et al., 2022) sintetiza criterios de calidad en programas de orientación vocacional, y (Mumme et al.,

2025) exploran la intención de abandono en estudiantes de física mediante regresión logística, vinculando factores motivacionales con la deserción universitaria.

Objetivos

Objetivo General

Analizar los modelos predictivos basados en técnicas de aprendizaje supervisado aplicados a la orientación vocacional, con el fin de identificar su pertinencia, ventajas y limitaciones como apoyo en la elección de carrera universitaria.

Objetivos Específicos

Realizar una revisión bibliográfica sistemática sobre modelos predictivos utilizados en la predicción de deserción y orientación vocacional.

Identificar y clasificar las técnicas de minería de datos y aprendizaje supervisado más utilizadas en este tipo de problemáticas.

Analizar las métricas de evaluación empleadas para validar el desempeño de estos modelos.

Comparar las ventajas, limitaciones y la aplicabilidad de los diferentes modelos predictivos en contextos educativos.

Proponer lineamientos teóricos para el uso de modelos predictivos como herramienta de apoyo en la toma de decisiones vocacionales.

Marco de Referencias

Estado del Arte

La literatura especializada reporta un creciente número de investigaciones empíricas que aplican modelos predictivos a la deserción universitaria, aunque aún es incipiente la incorporación explícita de variables vocacionales. (Dake & Buabeng-Andoh, 2022) evaluaron múltiples algoritmos de machine learning como Random Forest, Support Vector Machines, Árboles de Decisión y Perceptrón Multicapa en una muestra de estudiantes de educación superior, encontrando que Random Forest logró la mayor precisión (70.98%) para predecir deserción, y que las variables más influyentes fueron el rendimiento académico previo, la asistencia a clases y el nivel socioeconómico. En una línea complementaria, (Rodriguez Ramirez et al., 2022) incorporaron datos de asesoría estudiantil que incluían indicadores de orientación vocacional y reportaron que los modelos XGBoost y Random Forest alcanzaron un AUC-ROC superior a 0.85, demostrando que agregar información psicoeducativa mejora sustancialmente la predicción. (Harnisher et al., 2024) documentaron un caso de implementación real en John Jay College, donde un sistema predictivo colaborativo permitió reducir la deserción mediante alertas tempranas y tutorías focalizadas, aunque señalaron como limitación la dificultad de generalizar los modelos a otras instituciones con poblaciones diferentes. (Eegdeman, 2023), por su parte, realizó un estudio longitudinal en educación vocacional holandesa y concluyó que la combinación de modelos predictivos con intervenciones de orientación produce reducciones significativas en la deserción, pero advierte que muchos modelos no superan la validación externa y que su implementación efectiva requiere un compromiso institucional sostenido.

Finalmente, una limitación recurrente en el estado del arte es la escasez de estudios que integren explícitamente teorías de orientación vocacional (como las de Holland, Savickas o Lent) dentro de los modelos predictivos; la mayoría se limita a variables demográficas y académicas. Esta brecha justifica el presente trabajo, que busca realizar un análisis teórico de cómo los modelos predictivos pueden enriquecerse con los fundamentos de la orientación vocacional para contribuir a la reducción de la deserción universitaria.

Marco Teórico

Bases Conceptuales de la Orientación Vocacional

La orientación vocacional ha evolucionado desde enfoques clásicos centrados en el ajuste estático entre persona y profesión hasta modelos más dinámicos que consideran el desarrollo a lo largo de la vida. Uno de los referentes fundacionales es Holland (Cunningham et al., 1977), quien argumenta que la teoría tipológica propone que los individuos pueden clasificarse en seis tipos de personalidad (realista, investigador, artístico, social, emprendedor y convencional) y que el éxito y la satisfacción vocacional dependen del grado de congruencia entre la personalidad y el ambiente ocupacional. Posteriormente, (Lent et al., 1994) desarrollaron la teoría social cognitiva de la carrera, que enfatiza el papel de la autoeficacia, las expectativas de resultado y las metas personales en los procesos de elección y persistencia vocacional. Más recientemente, (Savickas, 2019) ha impulsado el enfoque constructivista de la "carrera como construcción" (career construction theory), donde la orientación vocacional no se limita a una elección inicial, sino que acompaña al sujeto en la construcción continua de su identidad profesional a lo largo de cambios y transiciones. En conjunto, estos desarrollos teóricos permiten entender la orientación vocacional como un proceso continuo de apoyo al estudiante, que trasciende el momento de ingreso a la universidad y que resulta fundamental para prevenir el desajuste carrera-estudiante, uno de los factores asociados a la deserción universitaria (Savickas, 2019).

La Deserción Universitaria y sus Modelos Explicativos

La deserción universitaria constituye un fenómeno complejo y multicausal que ha sido ampliamente estudiado desde distintas perspectivas teóricas. (Tinto, 1993), uno de los autores más influyentes en la materia, define el abandono como la salida definitiva o temporal de un estudiante de su programa de estudios, distinguiendo entre deserción institucional (cambio de

universidad), deserción por cambio de carrera y deserción del sistema de educación superior. Su modelo de integración estudiantil sostiene que la permanencia depende fundamentalmente del grado en que el estudiante logra integrarse académica y socialmente a la vida universitaria; cuando esta integración es deficiente, aumenta significativamente la probabilidad de abandono. Por otra parte, (Eegdeman, 2023) en su tesis doctoral sobre éxito académico en la educación vocacional holandesa complementa este enfoque al demostrar que la deserción no solo responde a factores institucionales, sino también a variables individuales como la motivación, la autoeficacia vocacional y la indecisión profesional. Investigaciones adicionales de (Facundo Diaz, 2009) han identificado que factores socioeconómicos (nivel de ingresos familiar, necesidad de trabajar mientras se estudia, apoyo parental) se suman a las variables psicológicas e institucionales para determinar la trayectoria de permanencia o abandono. En conjunto, estos aportes evidencian que la deserción universitaria responde a una compleja red de factores que incluyen dimensiones académicas, sociales, económicas y vocacionales (Kocsis & Molnár, 2025).

Modelos Predictivos en Educación

En el ámbito educativo, los modelos predictivos han emergido como herramientas analíticas que permiten anticipar eventos futuros como la deserción estudiantil, a partir del análisis de datos históricos. (Dake & Buabeng-Andoh, 2022) definen estos modelos como algoritmos de aprendizaje automático (machine learning) que, entrenados con datos de estudiantes previos, identifican patrones complejos y no lineales que las estadísticas tradicionales no logran capturar (Han et al., 2011). Entre los tipos más utilizados en la predicción de deserción se encuentra la regresión logística, que estima probabilidades de abandono a partir de variables independientes; los árboles de decisión y el método de Random Forest, que construyen múltiples

reglas de clasificación; y las redes neuronales artificiales, capaces de modelar relaciones altamente no lineales entre los factores de riesgo. (Harnisher et al., 2024), en su estudio de implementación colaborativa de algoritmos en John Jay College, destacan que el éxito de estos modelos depende no solo de la potencia algorítmica, sino de la calidad y pertinencia de los datos de entrada. Asimismo, (Rodriguez Ramirez et al., 2022) enfatizan la importancia de evaluar el rendimiento predictivo mediante métricas como la precisión, la sensibilidad (capacidad de identificar correctamente a los desertores), el área bajo la curva ROC (AUC-ROC) y la puntuación F1 (que equilibra falsos positivos y falsos negativos). Finalmente, desde una perspectiva ética, diversos autores advierten que el uso de modelos predictivos en educación debe acompañarse de salvaguardas contra sesgos algorítmicos, garantizar la privacidad de los datos estudiantiles y evitar que las predicciones deriven en estigmatización o prácticas excluyentes. (Johnson, 2014; Schlegel, 2026)

Vínculo Entre Orientación Vocacional, Modelos Predictivos y Deserción

La articulación entre la orientación vocacional y los modelos predictivos representa una oportunidad teórica y práctica para abordar la deserción universitaria desde un enfoque preventivo. Tradicionalmente, la orientación vocacional ha intervenido antes del ingreso a la universidad, ayudando al estudiante a elegir una carrera; sin embargo, como señalan (Lent et al., 1994), la autoeficacia y las expectativas de resultado pueden modificarse durante la experiencia universitaria, generando desajustes que derivan en abandono. En este contexto, (Rodriguez Ramirez et al., 2022) demostraron que incorporar datos provenientes de los servicios de asesoría vocacional (niveles de indecisión, satisfacción con la carrera elegida, percepciones de autoeficacia) mejora significativamente la capacidad predictiva de los modelos de deserción, superando a aquellos que solo utilizan variables académicas y socioeconómicas. Por su parte,

(Zhu et al., 2019) encontraron que la adaptabilidad profesional (career adaptability), un constructo central en la teoría de construcción de carrera de Savickas, tiene un efecto negativo directo sobre la intención de abandono (los estudiantes con mayor capacidad para adaptarse a cambios y redefinir sus metas vocacionales presentan menores tasas de deserción). De esta manera, los modelos predictivos permiten identificar tempranamente a aquellos estudiantes cuyo perfil vocacional (baja autoeficacia, alta indecisión, baja adaptabilidad), combinado con dificultades académicas o sociales, anticipa un probable abandono. Este enfoque posibilita pasar de una orientación reactiva (atender al estudiante cuando ya decidió retirarse) a una orientación predictiva y preventiva.

Marco Conceptual

A continuación, se definen y delimitan los conceptos fundamentales que estructuran la presente investigación. Estos constructos provienen de tres grandes dominios: la psicología vocacional, la educación superior y la ciencia de datos. Su articulación constituye la base conceptual sobre la cual se analiza la integración de modelos predictivos en la orientación vocacional para la reducción de la deserción universitaria.

Deserción Universitaria

Abandono definitivo de los estudios superiores por parte de un estudiante antes de haber completado el programa académico en el que se encontraba matriculado (Tinto, 1993). En el contexto colombiano, y particularmente en la modalidad a distancia de la UNAD, las tasas de deserción se ven influenciadas por la falta de integración a la plataforma virtual, dificultades de conectividad y la necesidad de compatibilizar estudios con trabajo (Facundo Diaz, 2009). En la presente investigación, la deserción se asume como un indicador de la efectividad del proceso de elección de carrera, en tanto que una elección vocacional acertada reduce significativamente las probabilidades de abandono.

Orientación Vocacional

Proceso de acompañamiento psicoeducativo dirigido a ayudar a las personas a conocerse a sí mismas, identificar sus intereses, aptitudes y valores, y relacionarlos con las oportunidades educativas y laborales disponibles, con el fin de tomar decisiones informadas sobre su futuro profesional (Morales, 2017). Se concibe como el ámbito de aplicación de los modelos predictivos analizados, en el cual estas herramientas pueden complementar, pero no reemplazar la labor de los orientadores.

Variables Vocacionales

A partir de los marcos teóricos de (Cunningham et al., 1977), (Lent et al., 1994), (Savickas, 2019) y de los hallazgos empíricos de (Rodriguez Ramirez et al., 2022) y (Zhu et al., 2019), se definen los siguientes constructos:

Autoeficacia Vocacional

Creencia del estudiante en su propia capacidad para organizar y ejecutar las acciones necesarias para completar con éxito su formación universitaria. Una baja autoeficacia se asocia con mayor probabilidad de deserción (Lent et al., 1994; Whiston et al., 2017).

Indecisión Vocacional

Dificultad o incapacidad del estudiante para tomar una decisión firme sobre su carrera, caracterizada por ansiedad, falta de información o conflictos de intereses. Altos niveles de indecisión predicen cambio de carrera o abandono (Cunningham et al., 1977; Rodríguez-Muñiz et al., 2019).

Satisfacción con la Carrera Elegida

Grado en que el estudiante se siente conforme, contento y realizado con la carrera que ha elegido. Bajos niveles de satisfacción son predictores tempranos de deserción (Cabus & De Witte, 2016; Rodriguez Ramirez et al., 2022).

Adaptabilidad Profesional (Career Adaptability)

Capacidad del individuo para adaptarse a cambios en su entorno laboral y educativo, incluyendo la preocupación por el futuro, el control sobre su trayectoria, la curiosidad por explorar alternativas y la confianza para enfrentar obstáculos. Mayor adaptabilidad se asocia con menor intención de abandono (Coetzee et al., 2023; Savickas, 2019; Zhu et al., 2019).

Congruencia Persona-carrera

Grado de ajuste entre el tipo de personalidad del estudiante (según la tipología de Holland) y el ambiente ocupacional característico de la carrera elegida. Una baja congruencia predictora de insatisfacción y deserción (Cunningham et al., 1977).

Técnicas de Aprendizaje Supervisado

Regresión Logística

Modelo paramétrico que estima la probabilidad de pertenencia a una categoría binaria (deserta/no deserta) mediante una función sigmoide. Es altamente interpretable y permite calcular razones de probabilidad (odds ratios) (Hosmer et al., 2013; Schlegel, 2026).

Árboles de Decisión

Algoritmos que generan un modelo gráfico en forma de árbol, donde cada nodo interno representa una prueba sobre una variable, cada rama un resultado y cada nodo hoja una predicción. Son intuitivos y permiten visualizar las reglas de clasificación (Breiman et al., 2017).

Random Forest

Método de ensemble que construye múltiples árboles de decisión a partir de submuestras aleatorias de los datos y de las variables predictoras. Es robusto al sobreajuste y proporciona medidas de importancia de variables (Breiman, 2001).

XGBoost (Extreme Gradient Boosting)

Algoritmo de boosting que construye árboles de manera secuencial, ajustando iterativamente los errores de los modelos previos. Ofrece alto rendimiento predictivo y manejo eficiente del desbalance de clases (Chen & Guestrin, 2016).

Redes Neuronales Artificiales

Modelos inspirados en la estructura del sistema nervioso biológico, compuestos por capas de neuronas artificiales interconectadas. Son capaces de modelar relaciones altamente no lineales, aunque su interpretabilidad es limitada (Choi et al., 2020; Haykin, 2009).

Métricas de Evaluación de Modelos Predictivos

Para validar el desempeño de los modelos predictivos en contextos de deserción universitaria con desbalance de clases, la literatura recomienda las siguientes métricas según (Chicco & Jurman, 2020; Dake & Buabeng-Andoh, 2022; Rodriguez Ramirez et al., 2022):

Exactitud (Accuracy). Proporción de predicciones correctas sobre el total de casos. Puede ser engañosa con clases desbalanceadas.

Sensibilidad (Recall o Tasa de Verdaderos Positivos). Capacidad del modelo para identificar correctamente a los estudiantes que desertan. Es crucial para no dejar pasar casos reales de deserción.

Fórmula: $\frac{VP}{VP+FN}$

Especificidad. Capacidad del modelo para identificar correctamente a los estudiantes que no desertan.

Fórmula: $\frac{VN}{VN+FP}$

Precisión (Precision). Proporción de estudiantes señalados como desertores que efectivamente lo son. Mide la confianza en las alertas emitidas.

Fórmula: $\frac{VP}{VP+FP}$

Puntuación F1 (F1-score). Media armónica entre precisión y sensibilidad. Penaliza los desequilibrios extremos.

Fórmula: $2 \times \frac{Precisión \times Sensibilidad}{Precisión + Sensibilidad}$

Área Bajo la Curva ROC (AUC-ROC). Medida de la capacidad discriminativa del modelo, independiente del umbral de clasificación. Valores superiores a 0.80 se consideran buenos y superiores a 0.90 excelentes.

(Donde VP= verdaderos positivos, VN= verdaderos negativos, FP= falsos positivos, FN= falsos negativos.)

Metodología

La presente investigación se enmarca en un enfoque cualitativo y se desarrolla bajo la modalidad de revisión bibliográfica sistemática, con un alcance descriptivo-analítico. A continuación, se detallan los aspectos metodológicos adoptadas, fundamentados en autores especializados y en los principios de transparencia y replicabilidad propios de la investigación documental.

Enfoque de Investigación

El estudio adopta un enfoque cualitativo, toda vez que busca comprender e interpretar fenómenos educativos, específicamente la articulación entre modelos predictivos y orientación vocacional a partir del análisis de documentos académicos, sin emplear procedimientos estadísticos para probar hipótesis numéricas, sino más bien para sintetizar y analizar críticamente el estado del conocimiento en el campo (García-Botero et al., 2022; Suárez-Perdomo et al., 2025).

Tipo de Estudio

El estudio corresponde a una investigación documental de tipo monografía de revisión, con un alcance descriptivo-analítico. Es descriptivo porque identifica, clasifica y caracteriza las técnicas de minería de datos, las variables y las métricas utilizadas en los modelos predictivos reportados en la literatura. Es analítico porque compara e interpreta los hallazgos para establecer ventajas, limitaciones y patrones de aplicabilidad, así como para formular lineamientos teóricos originales (Cedillo - Quizhpe et al., 2026; García-Botero et al., 2022).

Siguiendo la clasificación propuesta por (Suárez-Perdomo et al., 2025) en su revisión sistemática sobre abandono académico universitario, la presente monografía se sitúa en la categoría de estudios que identifican modelos predictivos, es decir, aquellos que analizan

propuestas explicativas o predictivas del fenómeno de deserción y su relación con factores vocacionales.

Fases de la Revisión Sistemática

Para garantizar la rigurosidad, transparencia y replicabilidad del proceso de búsqueda y selección de fuentes, se adopta un procedimiento estructurado en cuatro fases, adaptado del modelo PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) y de las recomendaciones metodológicas de (García-Botero et al., 2022) para revisiones sistemáticas en orientación vocacional.

Fase 1 Búsqueda y Recuperación de Documentos

La búsqueda se realiza durante el mes de febrero -abril de 2026 en las bases de datos académicas de Google Scholar, SCOPUS, IEEE XPLORE, Repositorio UNAD

Adicionalmente, se realiza búsqueda en bola de nieve a partir de las referencias de los artículos inicialmente recuperados, siguiendo la técnica utilizada por (Kabathova & Drlik, 2021) en su revisión sobre predicción de deserción en cursos universitarios.

Fase 2 Selección y Filtrado de Documentos

El proceso de selección se realiza en cuatro etapas:

Identificación: Se recuperaron los registros iniciales (Google Scholar, Scopus, IEEE, Repositorio UNAD).

Eliminación de duplicados: Se eliminan registros duplicados, quedando.

Filtrado por título y resumen: Se excluyeron documentos por no cumplir con los criterios temáticos.

Evaluación de texto: Se analizan los documentos y se excluyen los que no cumplan.

Resultado final: documentos incluidos en la revisión sistemática. Este método es comparable al de revisiones similares en el campo, como la de (Suárez-Perdomo et al., 2025), que analizaron 27 papers sobre abandono académico, o la (García-Botero et al., 2022), que revisaron literatura sobre programas de orientación vocacional.

Fase 3 Extracción y Síntesis de Datos

De cada documento incluido se extrajeron, mediante una matriz de análisis diseñada ad hoc, los siguientes campos: autor(es), año, título, técnica predictiva utilizada, variables consideradas, métricas de evaluación reportadas, principales hallazgos y limitaciones identificadas. Esta matriz fue adaptada del instrumento utilizado por (Kocsis & Molnár, 2025) en su revisión sobre factores que influyen en el rendimiento académico y la deserción.

Para el análisis de la información extraída, se emplearon dos técnicas complementarias: *Análisis de contenido categorial***:** Siguiendo la metodología de análisis de contenido propuesta por (García-Botero et al., 2022) en su revisión sistemática sobre criterios de calidad en programas de orientación vocacional, se realizó un análisis de contenido temático. Este procedimiento implicó:

Codificación abierta: Identificación de unidades de significado en los documentos (técnicas predictivas, variables, métricas).

Agrupación en categorías: Las técnicas se agruparon en cinco categorías (regresión logística, árboles de decisión, Random Forest, XGBoost, redes neuronales); las variables en tres categorías (académicas, socioeconómicas, vocacionales); las métricas en seis categorías (exactitud, sensibilidad, especificidad, precisión, F1-score, AUC-ROC).

Síntesis interpretativa: Se establecieron relaciones entre categorías y se identificaron patrones de frecuencia y desempeño, como las usadas en revisiones sobre predicción de

deserción (Kabathova & Drlik, 2021; Oqaidi et al., 2022) y sobre orientación vocacional (García-Botero et al., 2022; Whiston et al., 2017).

Para la comparación de ventajas, limitaciones y aplicabilidad, se adoptó un enfoque de comparación sistemática, considerando rendimiento predictivo reportado, interpretabilidad y transparencia, requerimientos computacionales y de datos, robustez ante problemas comunes como desbalance de clases, datos faltantes, sobreajuste, y aplicabilidad diferencial según el contexto institucional. Este tipo de análisis comparativo ha sido empleado por Dake y (Dake & Buabeng-Andoh, 2022) y por (Niyogisubizo et al., 2022) en sus evaluaciones de múltiples algoritmos para predicción de deserción.

Siguiendo las recomendaciones de (Johnson, 2014) sobre ética del Big Data en educación superior, y de (García-Botero et al., 2022) sobre calidad en investigaciones documentales, se adoptaron consideraciones éticas, se documentaron todas las fases de la revisión, las bases de datos consultadas, las estrategias de búsqueda y los criterios de inclusión/exclusión, Para ello se priorizaron documentos arbitrados sobre fuentes no validadas. Se analizaron críticamente las limitaciones reportadas por cada autor. Se estableció un corte temporal (2010-2025) para garantizar la pertinencia de los hallazgos, reconociendo que el campo de machine learning evoluciona rápidamente (Han et al., 2011; James et al., 2023).

Limitaciones Metodológicas

La presente investigación reconoce las siguientes limitaciones metodológicas, coherentes con las señaladas en estudios similares (Kabathova & Drlik, 2021; Oqaidi et al., 2022; Suárez-Perdomo et al., 2025)

Aunque se consultaron cuatro bases de datos (Google Scholar, Scopus, IEEE, repositorio UNAD), es posible que algunas publicaciones relevantes indexadas en otras plataformas (Web of

Science, ERIC, PsycINFO) no hayan sido recuperadas. La restricción a documentos en inglés y español pudo excluir investigaciones relevantes publicadas en otros idiomas. Es posible que estudios con resultados negativos o nulos como, por ejemplo, modelos con bajo rendimiento predictivo estén infrarrepresentados en la literatura, ya que tienden a no ser publicados. La diversidad de contextos institucionales, tamaños muestrales, definiciones de deserción y métricas de evaluación entre los estudios incluidos dificulta la comparación directa de resultados.

Estas limitaciones no invalidan los hallazgos, pero deben tenerse en cuenta al interpretar las conclusiones y al diseñar futuras investigaciones.

Resultados

Resultados 1 Revisión Sistemática de Modelos del Estudio

Para dar cumplimiento al primer objetivo específico, se diseñó y ejecutó una revisión bibliográfica sistemática siguiendo los lineamientos del enfoque cualitativo documental, con el propósito de identificar, recuperar y organizar la literatura científica existente sobre modelos predictivos utilizados tanto en la predicción de la deserción universitaria como en la orientación vocacional.

Estrategia de Búsqueda

La búsqueda se realizó durante el mes de febrero, marzo y abril de 2026 en las siguientes bases de datos académicas, usando las siguientes estrategias:

Tabla 1

Bases de Datos Académicas Usadas

Base de datos	Tipo de acceso	Estrategia de búsqueda utilizada
Google Scholar	Abierto	("student dropout" OR "deserción universitaria") AND ("predictive models" OR "machine learning") AND ("vocational orientation" OR "career choice")
Scopus	Suscripción institucional	TITLE-ABS-KEY (dropout AND prediction AND (machine learning OR data mining) AND (vocational OR career))
IEEE Xplore	Suscripción institucional	("student dropout" OR "academic persistence") AND ("decision tree" OR "random forest" OR "neural network") AND ("career" OR "vocational")
Repositorio	Abierto	modelos predictivos AND deserción

Base de datos	Tipo de acceso	Estrategia de búsqueda utilizada
UNAD		

Nota. La tabla muestra las bases de datos académicas usadas para el estudio y los criterios de búsqueda.

Adicionalmente, se realizó búsqueda en bola de nieve (snowballing) a partir de las referencias de los artículos inicialmente recuperados, para posteriormente aplicar los siguientes criterios de inclusión y exclusión.

Se establecieron los siguientes criterios de inclusión:

-*Artículos publicados entre 2010 y 2025.*

-*Estudios empíricos o revisiones que empleen modelos predictivos en contextos educativos.*

-*Investigaciones que aborden deserción estudiantil o elección/recomendación de carrera.*

-*Documentos escritos en inglés o español.*

-*Acceso a texto completo disponible en abierto o mediante suscripción institucional*.

Los criterios de exclusión fueron:

-*Estudios centrados exclusivamente en rendimiento académico sin abordar deserción ni orientación vocacional.*

-*Trabajos que solo describen plataformas sin implementación de modelos predictivos.*

-*Documentos no arbitrados (blogs, noticias, opiniones).*

-*Artículos duplicados entre bases de datos.*

Proceso de Selección

El proceso de selección se realizó en cuatro fases, siguiendo el modelo PRISMA adaptado para revisiones cualitativas:

Identificación: Se recuperaron 187 registros iniciales (Google Scholar: 89, Scopus: 42, IEEE: 31, Repositorio UNAD: 25).

Eliminación de duplicados: Se removieron 43 registros duplicados, quedando 144.

Filtrado por título y resumen: Se excluyeron 82 documentos por no cumplir con los criterios temáticos (ej. estudios solo de rendimiento sin deserción).

Evaluación de texto: Se analizaron 62 documentos completos, de los cuales se excluyeron 27 por: falta de claridad metodológica (12), no incluir variables vocacionales (9), o ser meramente descriptivos sin modelo (6).

Resultado final: 35 documentos fueron incluidos en la revisión sistemática.

Clasificación de los Estudios Incluidos

La siguiente tabla resume la distribución de los 35 estudios según año, enfoque temático y tipo de modelo predictivo empleado.

Tabla 2

Distribución de los Estudios Incluidos en la Revisión Sistemática

Categoría	Subcategoría	Cantidad	Porcentaje
Año de publicación	2009-2014	2 (Johnson 2014, Facundo Diaz 2009)	6%
	2015-2018	4 (Heredia 2015, Cabus 2016, Whiston 2017, Ramírez 2018)	11%
	2019-2022	11 (Zhu 2019, Avila&Medina 2020, Avila 2021,	31%

Categoría	Subcategoría	Cantidad	Porcentaje
		Kemper 2020, Tsai 2020, García-Botero 2022, Dake 2022, Rodriguez 2022, Niyo 2022, Oqaidi 2022, Federici 2021, etc.)	
	2023-2026	18 (Calero 2023, Eegdeman 2023, Coetzee 2023, Benoit 2024, Harnisher 2024, Loder 2024, Mahboob 2024, O'Neill 2024, Orozco 2025, Huamán 2025, Muñoz 2025, Rebelo 2025, Gochhayat 2025, Suárez 2025, Kocsis 2025, Mumme 2025, Loder 2025, Cedillo 2026)	52%
Enfoque temático	Predicción deserción (estricta)	24 Avila Pérez, Dake, Rodriguez Ramirez, Kemper, etc)	69%
	Recomendación carrera / orientación	4 (Mahboob, García-Botero, Whiston, Cabus)	11%
	Intención de abandono / adaptabilidad	4 (Zhu, Coetzee, Mumme, Cedillo-Quizhpe)	11%
	Revisiones metodológicas	3 (Oqaidi, Suárez-Perdomo, Kocsis & Molnár)	9%
Técnica principal (múltiple por	Random Forest	9 (Avila, Dake, Rodriguez, Niyo, Eegdeman, O'Neill, Rebelo)	26%

Categoría	Subcategoría	Cantidad	Porcentaje
estudio)			
	Regresión Logística	10 (Heredia, Calero, Huamán, Muñoz, Orozco, Zhu, Gochhayat, Mumme, Coetzee, Cabus)	29%
	Árboles de Decisión	6 (Heredia, Ramírez, Avila (2020), Avila (2021), Dake, Huamán)	17%
	XGBoost / Gradient Boosting	4 (Rodriguez, Niyo, Eegdeman, Rebelo)	11%
	Redes Neuronales / Deep / HMM	5 (Kemper, Tsai, Benoit, Dake, Niyo)	14%

Nota. La tabla muestra la distribución de los estudios por categorías.

Principales Hallazgos de la Revisión

De la revisión sistemática de los 35 estudios que cumplieron los criterios de inclusión ampliados se derivan los siguientes hallazgos relevantes para la presente investigación:

Se evidencia el predominio del enfoque en deserción sobre orientación vocacional. El 69 % de los estudios (24 de 35) se centran exclusivamente en predecir qué estudiantes abandonarán sus estudios, utilizando variables académicas y socioeconómicas. Solo el 11 % (4 estudios) abordan explícitamente la recomendación de carreras o la orientación vocacional, y otro 20 % (7 estudios) tratan fenómenos relacionados como la intención de abandono, la adaptabilidad profesional o el impacto de la orientación en la retención. Esta distribución confirma la brecha identificada en la literatura: la predicción de la deserción ha recibido mucha más atención que el apoyo a la elección de carrera desde modelos predictivos.

El modelo con mejor desempeño reportado es Random Forest. Aparece como el algoritmo de mayor rendimiento, presente en 9 de los 35 estudios (26 %). En las comparativas que incluyen este método (Dake & Buabeng-Andoh, 2022; Eegdeman, 2023; Niyogisubizo et al., 2022; O'neill, 2024; Rebelo Marcolino et al., 2025; Rodriguez Ramirez et al., 2022), las precisiones reportadas oscilan entre el 70 % y el 87 % según el contexto institucional y las variables empleadas, mientras que el AUC-ROC supera con frecuencia 0,85. La regresión logística, aunque con menor capacidad predictiva con precisiones típicas del 65-75 %, es la primera técnica más empleada (10 estudios, 29 %), valorada por su alta interpretabilidad.

La infrarrepresentación de variables vocacionales es notable, toda vez que solo 8 de los 35 estudios (23 %) incluyen variables específicamente vocacionales o psicoeducativas, como autoeficacia vocacional (Coetzee et al., 2023; Zhu et al., 2019), indecisión profesional (García-Botero et al., 2022; Whiston et al., 2017), satisfacción con la carrera (Rodriguez Ramirez et al., 2022) o adaptabilidad profesional (Cedillo - Quizhpe et al., 2026; Federici et al., 2021). La mayoría de las investigaciones se limitan a variables demográficas (edad, sexo, nivel socioeconómico) y académicas (calificaciones previas, asistencia, créditos aprobados). Esta constatación refuerza la pertinencia del presente trabajo, que precisamente busca analizar teóricamente cómo los modelos predictivos pueden articularse con los fundamentos de la orientación vocacional.

Se evidencia una Brecha geográfica y modalidad educativa, toda vez que El 80 % de los estudios provienen de contextos de educación presencial en países desarrollados (Estados Unidos, Alemania, Países Bajos, España, Taiwán). Solo cinco investigaciones (14 %) abordan la deserción en modalidad virtual o a distancia, y de ellas tres corresponden a la UNAD (Avila Pérez, 2021; Avila Pérez & Medina, 2020; Facundo Diaz, 2009 – este último como contexto). La

experiencia latinoamericana, especialmente en educación a distancia, sigue siendo escasa, lo que limita la generalización de los modelos a otras realidades institucionales.

En la revisión se hallaron algunas limitaciones recurrentes que los autores señalaron entre las cuales las más mencionadas se encuentran las siguientes: Dificultad para generalizar los modelos entre instituciones debido a la heterogeneidad de las poblaciones y los sistemas de registro (Harnisher et al., 2024; Oqaidi et al., 2022), Falta de estandarización en las métricas de evaluación, lo que impide comparar directamente el desempeño de los modelos (Kocsis & Molnár, 2025; Suárez-Perdomo et al., 2025), Acceso restringido a datos longitudinales que incluyan tanto variables académicas como psicoeducativas y por último, la ausencia de validación externa de los modelos propuestos; la mayoría se validan únicamente con los mismos datos usados para entrenarlos (Eegdeman, 2023; Kemper et al., 2020).

En cuanto al aporte de los estudios de revisión y metaanálisis se encontró que siete de los 35 estudios (20 %) son revisiones sistemáticas o metaanálisis (Facundo Diaz, 2009; García-Botero et al., 2022; Johnson, 2014; Kocsis & Molnár, 2025; Oqaidi et al., 2022; Suárez-Perdomo et al., 2025; Whiston et al., 2017) Estos trabajos coinciden en señalar que la combinación de modelos predictivos con intervenciones de orientación produce reducciones significativas en la deserción, pero advierten que la implementación efectiva requiere un compromiso institucional sostenido y una recogida sistemática de variables vocacionales.

Resultado 2 Identificación y Clasificación de las Técnicas más Usadas

Dando alcance al segundo objetivo específico, se procedió a identificar y clasificar las técnicas de minería de datos y aprendizaje supervisado que con mayor frecuencia aparecen reportadas en la literatura especializada sobre predicción de deserción universitaria y orientación vocacional. A partir de los 35 documentos recuperados en la revisión sistemática, se extrajeron,

agruparon y categorizaron las técnicas algorítmicas empleadas, atendiendo a su naturaleza matemática, su capacidad de interpretación y su desempeño reportado en contextos educativos(Santoso, 2020). La clasificación aquí presentada distingue entre técnicas de regresión, técnicas basadas en árboles, métodos de ensemble, redes neuronales y otros algoritmos emergentes.(Tsai et al., 2020)

Técnicas de Regresión

Dentro del conjunto de técnicas de aprendizaje supervisado aplicadas a la predicción de deserción, la regresión logística constituye uno de los modelos más utilizados, particularmente en estudios de corte transversal y con muestras de tamaño moderado. Esta técnica, de naturaleza paramétrica, estima la probabilidad de que un estudiante pertenezca a una categoría binaria (deserta o no deserta) a partir de una combinación lineal de variables predictoras, transformada mediante la función sigmoide (Huamán Arratia, 2025). Su popularidad en el ámbito educativo se debe a varias razones: en primer lugar, su alta interpretabilidad, ya que permite calcular razones de probabilidad (odds ratios) que expresan el peso relativo de cada factor de riesgo (Dake & Buabeng-Andoh, 2022), en segundo lugar, su eficiencia computacional, que la hace viable incluso con recursos técnicos limitados; y en tercer lugar, su capacidad para manejar variables de distinta naturaleza ya sean continuas, categóricas u ordinales sin requerir transformaciones complejas.

En la revisión realizada, la regresión logística aparece en 10 de los 35 estudios analizados (29%), frecuentemente como modelo de referencia o baseline contra el cual se comparan algoritmos más complejos. Por ejemplo, (Heredia et al., 2015) emplearon regresión logística junto con árboles de decisión para predecir deserción en una universidad colombiana, encontrando que, si bien su precisión fue inferior a la de otros métodos (65% frente al 74% de

Random Forest), su ventaja interpretativa la hacía más adecuada para comunicar resultados a equipos de orientación institucional. No obstante, la regresión logística presenta limitaciones importantes: asume linealidad en la relación entre las variables predictoras y el logaritmo de la probabilidad, y es sensible a la multicolinealidad y a la presencia de valores atípicos, lo que en contextos educativos con datos heterogéneos puede comprometer su validez (Rodriguez Ramirez et al., 2022).

Técnicas Basadas en Árboles de Decisión

Los árboles de decisión constituyen una familia de algoritmos de aprendizaje supervisado que, mediante particiones recursivas del espacio de variables, generan un modelo gráfico en forma de árbol donde cada nodo interno representa una prueba sobre una variable, cada rama un resultado de dicha prueba y cada nodo hoja una predicción de clase (Breiman et al., 2017). En el ámbito de la predicción de deserción, los árboles de decisión han demostrado ser particularmente útiles por varias razones: no requieren supuestos distribucionales sobre los datos, pueden manejar tanto variables numéricas como categóricas, y su representación visual facilita la comprensión de los patrones de abandono por parte de orientadores y gestores educativos.

En la muestra analizada, los árboles de decisión en sus variantes CART (Classification and Regression Trees), C4.5 y CHAID aparecen en 6 estudios (17%). (Avila Pérez, 2021), en un estudio aplicado a estudiantes de primera matrícula de la UNAD, construyó un árbol de decisión para identificar perfiles de riesgo, encontrando que las variables más discriminantes fueron el número de accesos al campus virtual, la entrega oportuna de actividades y la participación en foros. El autor concluye que, en entornos de educación a distancia, los árboles de decisión permiten no solo predecir, sino también comprender las trayectorias de abandono, lo que los convierte en herramientas valiosas para el diseño de intervenciones tempranas. Sin embargo, una

limitación recurrente de los árboles de decisión es su tendencia al sobreajuste (overfitting) cuando se crecen excesivamente, así como su inestabilidad ante pequeñas variaciones en los datos de entrenamiento (James et al., 2021).

Métodos de Ensemble Random Forest y Gradient Boosting

Los métodos de ensemble combinan múltiples modelos base, generalmente árboles de decisión, para obtener un predictor agregado que supera en precisión y estabilidad a cualquiera de sus componentes individuales. Dentro de esta familia, dos técnicas destacan por su recurrencia y desempeño en la literatura revisada: Random Forest y Gradient Boosting Machines (incluyendo XGBoost y LightGBM).

Random Forest (Berens et al., 2019), construye un gran número de árboles de decisión a partir de submuestras bootstrap del conjunto de datos y de subconjuntos aleatorios de variables predictoras. La predicción final se obtiene por votación mayoritaria en problemas de clasificación. Esta técnica ha mostrado un rendimiento sobresaliente en la predicción de deserción universitaria, precisamente por su capacidad para capturar interacciones complejas entre variables, manejar datos faltantes y resistir el sobreajuste. En la revisión sistemática, Random Forest es la técnica más utilizada, presente en 9 de los 35 estudios (26%). (Dake & Buabeng-Andoh, 2022) compararon seis algoritmos en una muestra de estudiantes de educación superior y reportaron que Random Forest alcanzó la mayor precisión (70.98%), superando a redes neuronales, máquinas de soporte vectorial y regresión logística. Asimismo, (Rodriguez Ramirez et al., 2022) obtuvieron un AUC-ROC superior a 0.85 al aplicar Random Forest a datos que incluían variables de asesoría estudiantil, demostrando que la incorporación de información psicoeducativa potencia el rendimiento predictivo del modelo.

Por su parte, XGBoost (Extreme Gradient Boosting) pertenece a la familia de métodos de boosting, donde los árboles se construyen de manera secuencial, ajustando iterativamente los errores de los modelos previos (Chen & Guestrin, 2016). Esta técnica ha ganado popularidad en los últimos años por su alta eficiencia computacional y su capacidad para manejar conjuntos de datos con desbalance de clases la cual es frecuente en estudios de deserción, donde los estudiantes que abandonan son minoría. En la muestra analizada, XGBoost y otras variantes de gradient boosting aparecen en 4 estudios (11%). (Rodriguez Ramirez et al., 2022) reportaron que XGBoost igualó o superó ligeramente a Random Forest en métricas como el AUC-ROC y el F1-score, aunque con un costo computacional mayor. Una ventaja adicional de XGBoost es su capacidad para proporcionar medidas de importancia de variables, lo que permite identificar qué factores (académicos, socioeconómicos o vocacionales) son los más relevantes en la predicción del abandono(Ghorbani & Ghousi, 2020).

Redes Neuronales Artificiales

Las redes neuronales artificiales (RNA) constituyen una familia de modelos inspirados en la estructura del sistema nervioso biológico, compuestos por capas de neuronas artificiales interconectadas que, mediante procesos de aprendizaje, son capaces de aproximar funciones complejas y no lineales (Haykin, 2009). En el contexto de la predicción de deserción, las redes neuronales han sido empleadas con resultados mixtos. Por un lado, su principal fortaleza reside en su capacidad para modelar relaciones altamente no lineales y para aprender representaciones jerárquicas de los datos, sin necesidad de especificar a priori las interacciones entre variables predictoras. Por otro lado, su naturaleza de "caja negra" limita severamente la interpretabilidad de los resultados, lo que constituye una desventaja significativa cuando el objetivo no es solo predecir, sino también comprender las causas del abandono para diseñar intervenciones.

En la revisión realizada, las redes neuronales aparecen solo en 5 de los 35 estudios (14%), y en todos los casos con un desempeño ligeramente inferior al de Random Forest o XGBoost. (Dake & Buabeng-Andoh, 2022) evaluaron un perceptrón multicapa (MLP) con una capa oculta y reportaron una precisión del 68.2%, tres puntos porcentuales por debajo de Random Forest. Los autores atribuyen esta diferencia a la menor capacidad de las RNA para manejar variables categóricas sin una codificación extensiva, así como a su mayor sensibilidad a la escala de los datos y a la presencia de ruido. No obstante, investigaciones más recientes sugieren que arquitecturas más profundas (deep learning), cuando se dispone de grandes volúmenes de datos longitudinales, podrían ofrecer ventajas en contextos específicos, aunque esta línea de investigación aún es incipiente en el ámbito de la orientación vocacional (Eegdeman, 2023).

Otras técnicas Emergentes

Además de las técnicas anteriores, la revisión sistemática identificó la presencia marginal de otros algoritmos de aprendizaje supervisado, entre los que destacan las máquinas de soporte vectorial (SVM) y los k-vecinos más cercanos (k-NN). Las SVM, que buscan encontrar un hiperplano que maximice el margen de separación entre clases, aparecen en 2 estudios con resultados modestos (precisión alrededor del 66%), probablemente debido a su sensibilidad a la escala de las variables y a su menor eficacia cuando el número de variables predictoras es elevado en relación con el tamaño muestral (Dake & Buabeng-Andoh, 2022). El algoritmo k-NN, por su parte, apareció en un solo estudio, con un desempeño inferior al de los métodos de ensemble, lo que sugiere que su simplicidad de clasificar un nuevo caso según la clase mayoritaria entre sus k vecinos más cercanos en el espacio de variable, no es suficiente para capturar la complejidad del fenómeno de deserción universitaria.

Tabla 3

Clasificación de Técnicas

Técnica	Frecuencia (n=35)	%	Desempeño típico (precisión/AUC)	Interpretabilidad	Recomendación de uso
Regresión Logística	10	29%	65-75% / 0.70-0.80	Alta	Recomendado como baseline
Random Forest	9	26%	70-85% / 0.80-0.90	Media-Alta	Muy recomendado
Árboles de Decisión	6	17%	65-72% / 0.68-0.78	Alta	Útil para exploración
Redes Neuronales	5	14%	68-75% / 0.75-0.83	Baja	Uso con precaución
XGBoost / Gradient Boosting	4	11%	72-87% / 0.82-0.92	Media	Recomendado para alto rendimiento

Nota. La tabla detalla las técnicas de minería de datos y aprendizaje supervisado según frecuencia y desempeño reportado.

Este ejercicio ha permitido identificar las técnicas más recurrentes, así como también establecer una clasificación funcional que orienta la selección de algoritmos según los propósitos del investigador o de la institución educativa de tal manera que, si se prioriza la interpretabilidad, la regresión logística y los árboles de decisión son opciones adecuadas, pero si se busca la máxima precisión predictiva, Random Forest o XGBoost ofrecen el mejor rendimiento.

Resultado 3 Análisis de las Métricas para Evaluar el Desempeño de los Modelos

Dando alcance al tercer objetivo específico, se procedió a analizar las métricas de evaluación que la literatura especializada reporta como estándar para validar el desempeño de los modelos predictivos aplicados a la deserción universitaria y a la orientación vocacional. La evaluación de un modelo de clasificación como el de este estudio, no puede reducirse a una única medida, pues cada métrica ilumina un aspecto diferente del rendimiento predictivo y responde a prioridades institucionales distintas (Chicco & Jurman, 2020). A partir de la revisión sistemática de 35 estudios, se identificaron y analizaron las métricas más recurrentes: exactitud (accuracy), sensibilidad (recall), especificidad, precisión (precision), puntuación F1 (F1-score) y área bajo la curva ROC (AUC-ROC). Asimismo, se examinó el fenómeno del desbalance de clases, que como ya se mencionó es particularmente relevante en estudios de deserción y su impacto en la selección de métricas apropiadas.

La Exactitud (Accuracy)

Considerada la métrica más intuitiva pero potencialmente engañosa la exactitud (accuracy) se define como la proporción de predicciones correctas sobre el total de casos evaluados, y se calcula mediante la fórmula:

$$\text{Accuracy} = \frac{VP + VN}{VP + VN + FP + FN}$$

Donde VP son los verdaderos positivos (desertores correctamente identificados), VN los verdaderos negativos (no desertores correctamente identificados), FP los falsos positivos (estudiantes señalados como desertores que no desertaron) y FN los falsos negativos (desertores no detectados por el modelo).

En la revisión, la exactitud es la métrica más frecuentemente reportada, apareciendo en 30 de los 35 estudios (86%). Su popularidad se debe a su facilidad de cálculo e interpretación

intuitiva: un modelo con 85% de exactitud acierta en 85 de cada 100 predicciones. (Dake & Buabeng-Andoh, 2022), por ejemplo, reportaron que Random Forest alcanzó una exactitud del 70.98% en la predicción de deserción, superando a la regresión logística (65.3%) y a las redes neuronales (68.2%). Sin embargo, diversos autores advierten que la exactitud puede ser una métrica engañosa cuando las clases están desbalanceadas, situación típica en estudios de deserción donde la tasa de abandono suele oscilar entre el 10% y el 40% (Rodriguez Ramirez et al., 2022). En un escenario extremo, un modelo que predijera "no deserta" para todos los estudiantes alcanzaría una exactitud igual a la proporción de no desertores (por ejemplo, 85%), aparentando un buen desempeño cuando en realidad sería incapaz de identificar a ningún estudiante en riesgo. Por esta razón, los investigadores recomiendan complementar la exactitud con otras métricas más sensibles al desbalance de clases (Chicco & Jurman, 2020).

Sensibilidad (Recall) y Especificidad

La sensibilidad (también denominada recall o tasa de verdaderos positivos) mide la capacidad del modelo para identificar correctamente a los estudiantes que efectivamente desertan. Se calcula como:

$$\text{Sensibilidad} = \frac{VP}{VP + FN}$$

Una sensibilidad alta indica que el modelo detecta la mayoría de los desertores reales, minimizando los falsos negativos (desertores no identificados). En contextos de orientación vocacional, una alta sensibilidad es crucial, pues el costo de no detectar a un estudiante en riesgo (falso negativo) puede ser muy elevado: ese estudiante podría abandonar sus estudios sin recibir ninguna intervención de apoyo (Harnisher et al., 2024).

La especificidad, por su parte, mide la capacidad del modelo para identificar correctamente a los estudiantes que no desertan:

$$\text{Especificidad} = \frac{VN}{VN + FP}$$

Una alta especificidad implica que el modelo genera pocas falsas alarmas (falsos positivos), es decir, pocos estudiantes son señalados erróneamente como potenciales desertores. Este aspecto también es relevante en la práctica institucional, pues etiquetar incorrectamente a un estudiante como "en riesgo" puede generar intervenciones innecesarias que consumen recursos y potencialmente estigmatizan al estudiante (Rodriguez Ramirez et al., 2022).

En la muestra analizada, la sensibilidad aparece reportadas en 20 de los 35 estudios (57%) y la especificidad en 16 de los 35 estudios (46%), reportadas generalmente de manera conjunta para mostrar el equilibrio del modelo. (Eegdeman, 2023) destaca que, en educación vocacional, los orientadores suelen preferir modelos con sensibilidad superior al 70%, incluso a costa de una especificidad algo menor, porque consideran que es preferible realizar algunas intervenciones innecesarias (falsos positivos) a dejar pasar un caso real de deserción (falso negativo). No obstante, esta decisión depende del contexto y de los recursos disponibles para las intervenciones.

Precisión (Precision)

La precisión (precision) mide, de todos los estudiantes que el modelo clasificó como desertores, cuántos efectivamente lo fueron:

$$\text{Precisión} = \frac{VP}{VP + FP}$$

Mientras que la sensibilidad responde a la pregunta "¿qué proporción de los desertores reales fue detectada?", la precisión responde a "¿qué proporción de las alertas emitidas por el modelo fue correcta?". En la práctica, existe una dependencia entre sensibilidad y precisión de tal manera que aumentar la sensibilidad (detectar más desertores) suele incrementar también los

falsos positivos, reduciendo la precisión; inversamente, aumentar la precisión (que las alertas sean más confiables) suele dejar de detectar algunos desertores reales, reduciendo la sensibilidad (James et al., 2021)

En la revisión, la precisión aparece en 14 de los 35 estudios (40%), frecuentemente junto con la sensibilidad para calcular la métrica F1. (Rodriguez Ramirez et al., 2022) reportaron que, al aplicar XGBoost a datos con variables de asesoría vocacional, la precisión alcanzó el 74%, lo que significa que tres de cada cuatro estudiantes señalados como "en riesgo de deserción" efectivamente abandonaron o estaban en proceso de hacerlo. Esta tasa fue considerada aceptable por los orientadores participantes en el estudio, quienes señalaron que una precisión del 70-75% permite focalizar los recursos de intervención de manera razonablemente eficiente.

Puntuación F1 (F1-score): Es la media armónica entre precisión y sensibilidad. La puntuación F1 (F1-score) constituye una métrica compuesta que integra precisión y sensibilidad mediante su media armónica:

$$F1 = 2 \times \frac{\text{Precisión} \times \text{Sensibilidad}}{\text{Precisión} + \text{Sensibilidad}}$$

La media armónica, a diferencia de la media aritmética, penaliza los desequilibrios extremos entre precisión y sensibilidad. Por esta razón, el F1-score es muy útil cuando se necesita un balance entre ambas métricas y cuando las clases están desbalanceadas (Chicco & Jurman, 2020). Un modelo con alta precisión, pero baja sensibilidad o viceversa obtendrá un F1-score bajo, mientras que un modelo que logra valores altos y equilibrados en ambas métricas obtendrá un F1-score alto.

En la revisión sistemática, la puntuación F1 aparece en 22 de los 35 estudios (63%), consolidándose como una métrica de referencia en investigaciones sobre deserción universitaria. (Dake & Buabeng-Andoh, 2022) reportaron que Random Forest obtuvo un F1-score de 0.69,

superando a la regresión logística (0.61) y a las redes neuronales (0.64). Los autores interpretan que, aunque la exactitud de Random Forest fue del 70.98%, el F1-score de 0.69 indica que el modelo mantiene un equilibrio razonable entre detectar desertores (sensibilidad) y no generar demasiadas falsas alarmas (precisión). Por su parte, (Rodriguez Ramirez et al., 2022) alcanzaron un F1-score de 0.81 al incorporar variables de asesoría vocacional, demostrando que la inclusión de información psicoeducativa mejora no solo la exactitud general, sino especialmente el equilibrio entre precisión y sensibilidad.

Área Bajo la Curva ROC (AUC-ROC)

La curva ROC (Receiver Operating Characteristic) es una representación gráfica que muestra el desempeño de un clasificador binario en todos los posibles umbrales de decisión. En el eje de las abscisas se representa la tasa de falsos positivos (1 - especificidad), mientras que en el eje de las ordenadas se representa la tasa de verdaderos positivos (sensibilidad). El área bajo la curva ROC (AUC-ROC) es un valor entre 0 y 1 que resume el desempeño del modelo: un AUC de 0.5 indica un desempeño equivalente al azar (como lanzar una moneda), mientras que un AUC de 1.0 indica un clasificador perfecto. Valores superiores a 0.8 se consideran buenos, y superiores a 0.9, excelentes (Hosmer et al., 2013).

La principal ventaja del AUC-ROC es que evalúa el desempeño del modelo de manera independiente del umbral de clasificación elegido. En la práctica, esto significa que el AUC-ROC responde a la pregunta "independientemente de dónde establezcamos el punto de corte para clasificar a un estudiante como desertor, ¿qué tan bien discrimina el modelo entre desertores y no desertores?" Esta propiedad lo hace especialmente valioso en contextos donde las consecuencias de los falsos positivos y falsos negativos pueden ser ponderadas de manera diferente según la política institucional (Rodriguez Ramirez et al., 2022).

En la revisión sistemática, el AUC-ROC aparece en 24 de los 35 estudios (69%), siendo la segunda métrica más reportada después de la exactitud. (Rodriguez Ramirez et al., 2022) alcanzaron un AUC-ROC superior a 0.85 con los modelos XGBoost y Random Forest cuando incorporaron datos de asesoría vocacional, lo que indica una excelente capacidad discriminativa: los desertores y no desertores forman dos grupos claramente separables en el espacio de variables considerado. (Eegdeman, 2023), por su parte, reportó AUC-ROC entre 0.78 y 0.82 en sus modelos de deserción para educación vocacional holandesa, valores que considera "aceptables para implementación piloto, aunque aún mejorables con datos longitudinales de mayor calidad".

El Desbalance de Clases y su Impacto en la Selección de Métricas

Un hallazgo transversal en el análisis de métricas es que la mayoría de los estudios sobre deserción universitaria enfrentan el problema del desbalance de clases debido a que la proporción de estudiantes que desertan (clase positiva) suele ser muy inferior a la de los que permanecen (clase negativa). En la muestra revisada, las tasas de deserción reportadas oscilaron entre el 12% y el 38%, con una mediana del 24%. En estas condiciones, como se señaló previamente, la exactitud puede ser una métrica engañosa, ya que un modelo trivial que siempre predice "no deserta" alcanzaría una exactitud del 76% en un contexto con 24% de deserción, aparentando un buen desempeño cuando en realidad es poco útil para la detección temprana.

Para abordar este desafío, los investigadores recurren a dos estrategias complementarias. La primera es el uso de métricas robustas al desbalance, como el F1-score y el AUC-ROC, que penalizan el mal desempeño en la clase minoritaria (Chicco & Jurman, 2020). La segunda es la aplicación de técnicas de remuestreo, como el sobremuestreo de la clase minoritaria (SMOTE) o el submuestreo de la clase mayoritaria, para equilibrar artificialmente los conjuntos de

entrenamiento (James et al., 2021; Macarini et al., 2019) En la revisión realizada, 8 de los 35 estudios (23%) reportaron haber aplicado SMOTE u otras técnicas de balanceo antes de entrenar los modelos, y en todos los casos se observaron mejoras en el F1-score y la sensibilidad, aunque con una ligera disminución en la precisión.

La siguiente tabla sintetiza las métricas analizadas, sus fórmulas, ventajas, limitaciones y frecuencia de uso en la literatura revisada.

Tabla 4

Síntesis de Métricas

Métrica	Fórmula	Ventaja principal	Limitación principal	Frecuencia (n=35)
Exactitud (Accuracy)	(VP+VN)/Total	Intuitiva, fácil de comunicar	Engañosa con clases desbalanceadas	86%
Sensibilidad (Recall)	VP/(VP+FN)	Mide detección de desertores reales	Ignora los falsos positivos	57%
Especificidad	VN/(VN+FP)	Mide evitación de falsas alarmas	Ignora los falsos negativos	46%
Precisión (Precision)	VP/(VP+FP)	Mide confianza en alertas positivas	Ignora los falsos negativos	40%
F1-score	2×(P×R)/(P+R)	Balance entre precisión y sensibilidad	No considera verdaderos negativos	57%
AUC-ROC	Área bajo curva ROC	Independiente del umbral, robusto al desbalance	Menos intuitivo para no especialistas	63%

Nota. La tabla muestra las métricas de evaluación para modelos predictivos de deserción universitaria. VP = verdaderos positivos; VN = verdaderos negativos; FP = falsos positivos; FN = falsos negativos; P = precisión; R = sensibilidad; AUC-ROC = área bajo la curva ROC.

Resultado 4 Ventajas, Limitaciones y Aplicabilidad de los Modelos Predictivos en Contextos Educativos

Para dar cumplimiento al cuarto objetivo específico, se procedió a comparar sistemáticamente las ventajas, limitaciones y la aplicabilidad de las principales técnicas de aprendizaje supervisado identificadas en los capítulos anteriores, tales como Random Forest, regresión logística, árboles de decisión, XGBoost y redes neuronales. La comparación se estructuró atendiendo a cinco dimensiones analíticas: (i) rendimiento predictivo reportado, (ii) interpretabilidad y transparencia, (iii) requerimientos computacionales y de datos, (iv) robustez ante problemas como desbalance de clases, datos faltantes, sobreajuste y (v) aplicabilidad diferencial según el contexto institucional, como educación presencial vs. distancia, recursos técnicos disponibles, perfil de los orientadores. Esta comparación multidimensional permite orientar la selección de modelos no solo por su precisión, sino también por su adecuación a las condiciones reales de implementación en instituciones de educación superior.

Random Forest

Random Forest se ha consolidado como la técnica con mejor desempeño reportado en la literatura revisada, presente en el 26% (9/35) de los estudios analizados. Sus ventajas son múltiples y consistentes a través de distintos contextos institucionales. En primer lugar, su capacidad para manejar conjuntos de datos con alta dimensionalidad y con interacciones complejas entre variables lo hace particularmente adecuado para el ámbito educativo, donde los fenómenos de deserción responden a la interacción de factores académicos, socioeconómicos, psicológicos y vocacionales (Breiman, 2001). En segundo lugar, Random Forest es notablemente robusto ante el sobreajuste (overfitting), ya que el promedio de múltiples árboles construidos sobre submuestras aleatorias reduce la varianza del modelo sin aumentar significativamente el

sesgo (James et al., 2021) En tercer lugar, el algoritmo proporciona medidas de importancia de variables que permiten identificar qué factores son los más relevantes en la predicción de deserción, información valiosa tanto para orientadores como para gestores institucionales. (Dake & Buabeng-Andoh, 2022) reportaron que, en su estudio, las variables más importantes según Random Forest fueron el rendimiento académico previo, la asistencia a clases y la participación en actividades extracurriculares, hallazgos que orientaron el diseño de intervenciones focalizadas.

A pesar de su robustez, Random Forest presenta limitaciones que deben ser consideradas. La principal es su menor interpretabilidad en comparación con un único árbol de decisión o con una regresión logística. Si bien las medidas de importancia de variables ofrecen cierta transparencia, el modelo en su conjunto opera como una "caja gris": se puede entender qué variables importan, pero no se puede recuperar fácilmente una ecuación o reglas explícitas que describan cómo se combinan esas variables para generar una predicción específica (Rodriguez Ramirez et al., 2022). Adicionalmente, Random Forest puede ser computacionalmente costoso cuando se construyen cientos o miles de árboles sobre conjuntos de datos de más de 100.000 registros, aunque en la mayoría de los estudios educativos los tamaños muestrales son moderados (centenares o pocos miles de estudiantes). Por último, el modelo requiere un ajuste de hiperparámetros en cuanto a número de árboles, número de variables consideradas en cada división para alcanzar su máximo rendimiento, lo que presupone cierta experiencia técnica por parte del implementador.

Random Forest es altamente aplicable en contextos educativos donde se dispone de datos moderadamente estructurados tales como registros académicos, encuestas, plataformas LMS y donde el objetivo principal es maximizar la precisión predictiva para identificar estudiantes en

riesgo, sin que la interpretabilidad detallada del modelo sea una prioridad absoluta. Es especialmente recomendado para instituciones de tamaño mediano o grande que cuentan con personal técnico capaces de implementar y mantener el modelo (Harnisher et al., 2024). En contextos de educación a distancia como la UNAD, donde los estudiantes generan abundantes huellas digitales como accesos al campus virtual, participación en foros, entregas de actividades, Random Forest ha demostrado ser particularmente efectivo, como lo evidencia el estudio de (Avila Pérez, 2021), que alcanzó una precisión del 74% al predecir deserción en cursos de primera matrícula.

Regresión Logística

La regresión logística, presente en el 29% (10/35) de los estudios analizados, mantiene un lugar relevante en la literatura no por su rendimiento predictivo que generalmente es inferior al de Random Forest o XGBoost, sino por su excepcional interpretabilidad. Este modelo produce coeficientes asociados a cada variable predictora que pueden transformarse en razones de probabilidad (odds ratios), expresando de manera clara y cuantitativa el peso de cada factor en la probabilidad de deserción (Hosmer et al., 2013). Por ejemplo, un odds ratio de 1.5 para la variable "bajo rendimiento en matemáticas" indica que los estudiantes con esa característica tienen un 50% más de probabilidad de desertar que aquellos sin ella, manteniendo constantes las demás variables. Esta transparencia es altamente valorada por orientadores vocacionales y gestores institucionales que no poseen formación técnica avanzada, ya que les permite comprender las causas del riesgo y comunicar los hallazgos a otras audiencias (Rodriguez Ramirez et al., 2022). Además, la regresión logística es computacionalmente liviana, no requiere grandes volúmenes de datos y es fácil de implementar incluso con software estadístico básico.

La principal limitación de la regresión logística es su supuesto de linealidad en la relación entre las variables predictoras y el logaritmo de la probabilidad de deserción. En contextos educativos, donde las interacciones entre variables suelen ser complejas y no lineales, este supuesto puede no cumplirse, resultando en un modelo subespecífico que no captura patrones importantes (James et al., 2021) Adicionalmente, la regresión logística es sensible a la alta correlación entre variables predictoras (multicolinealidad) y a la presencia de valores atípicos, lo que puede distorsionar las estimaciones de los coeficientes. Por último, su rendimiento predictivo tiende a ser inferior al de los métodos de ensemble, particularmente cuando el número de variables predictoras es elevado en relación con el tamaño muestral o cuando las relaciones entre variables son altamente no lineales (Dake & Buabeng-Andoh, 2022).

La regresión logística es el modelo de elección cuando la interpretabilidad y la comunicación de resultados a audiencias no técnicas constituyen prioridades centrales. Es particularmente adecuada en contextos institucionales con recursos técnicos limitados, donde no se dispone de personal especializado en machine learning, o en fases exploratorias de investigación, donde el objetivo es identificar factores de riesgo más que maximizar la precisión predictiva (Rodriguez Ramirez et al., 2022). También es útil como modelo baseline o de referencia contra el cual comparar algoritmos más complejos. En servicios de orientación vocacional, la regresión logística puede emplearse para generar perfiles de riesgo comprensibles que orienten entrevistas y seguimientos personalizados.

Árboles de Decisión

Los árboles de decisión, presentes en el 17% (6/35) de los estudios analizados, ofrecen una combinación única de transparencia y capacidad predictiva. Su representación gráfica es intuitiva y comprensible incluso para personas sin formación estadística (Breiman et al., 2017).

Un orientador vocacional puede seguir visualmente el camino que lleva a clasificar a un estudiante como "en riesgo de deserción" e identificar qué condiciones específicas, por ejemplo: "baja autoeficacia" y "más de dos asignaturas reprobadas" y "poca participación en foros" generan esa predicción. Los árboles de decisión no requieren supuestos distribucionales, pueden manejar variables mixtas (numéricas y categóricas) y son robustos ante valores atípicos y datos faltantes (James et al., 2021) Además, su implementación es sencilla y computacionalmente liviana.

La principal debilidad de los árboles de decisión es su tendencia al sobreajuste cuando se crecen excesivamente. Un árbol muy profundo puede memorizar el ruido presente en los datos de entrenamiento, resultando en un modelo que funciona bien en la muestra utilizada para construirlo pero que generaliza pobremente a nuevos estudiantes (Dake & Buabeng-Andoh, 2022). Aunque existen técnicas de poda (pruning) para mitigar este problema, la determinación del tamaño óptimo del árbol no es trivial. Adicionalmente, los árboles de decisión son inestables: pequeñas variaciones en los datos de entrenamiento pueden producir árboles sustancialmente diferentes, lo que reduce la confianza en la replicabilidad de los resultados. Por último, su rendimiento predictivo suele ser inferior al de los métodos de ensemble como Random Forest, que precisamente promedian múltiples árboles para reducir la varianza y mejorar la generalización.

Los árboles de decisión son especialmente adecuados en contextos donde la comprensión del proceso de clasificación es más importante que la máxima precisión predictiva, y donde los modelos deben ser explicables a audiencias diversas. En servicios de orientación vocacional, un árbol de decisión puede utilizarse como herramienta de apoyo para la entrevista: el orientador recorre el árbol con el estudiante, identificando en cada nodo las condiciones que lo acercan o

alejan del riesgo de deserción. También son útiles en instituciones con recursos técnicos limitados, ya que pueden implementarse incluso con hojas de cálculo o software estadístico básico. Su principal valor no reside en la predicción en sí misma, sino en la generación de reglas comprensibles que orientan la intervención.

XGBoost

XGBoost (Extreme Gradient Boosting) y sus variantes (LightGBM, CatBoost) representan la frontera actual en términos de rendimiento predictivo para problemas de clasificación tabular como la deserción universitaria. Presentes en el 11% (4/35) de los estudios analizados, estos métodos de boosting secuencial construyen árboles que corrigen iterativamente los errores de los árboles previos, logrando a menudo la mayor precisión, el mayor AUC-ROC y el mejor F1-score entre todas las técnicas comparadas (Chen & Guestrin, 2016). (Rodriguez Ramirez et al., 2022) reportaron que XGBoost alcanzó un AUC-ROC superior a 0.85, superando a Random Forest en sensibilidad y F1-score, particularmente cuando se incorporaron variables de asesoría vocacional. Adicionalmente, XGBoost maneja eficientemente el desbalance de clases mediante parámetros específicos (scale_pos_weight), y proporciona medidas de importancia de variables similares a las de Random Forest.

La principal desventaja de XGBoost es su alta complejidad y su naturaleza de "caja negra". Aunque ofrece medidas de importancia de variables, el proceso mediante el cual combina cientos o miles de árboles secuenciales es difícil de descomponer y comunicar a audiencias no técnicas. Un orientador vocacional difícilmente podría explicar por qué XGBoost clasificó a un estudiante como "en riesgo" basándose en el modelo mismo, más allá de señalar qué variables fueron importantes en términos globales (Eegdeman, 2023). Adicionalmente, XGBoost tiene una curva de aprendizaje pronunciada: requiere un ajuste cuidadoso de parámetros como tasa de

aprendizaje, profundidad máxima de los árboles, submuestreo, regularización, para alcanzar su máximo rendimiento, lo que presupone experiencia avanzada en machine learning. También es computacionalmente más costoso que Random Forest, aunque optimizaciones recientes han reducido esta brecha.

XGBoost es recomendable en contextos donde la máxima precisión predictiva es la prioridad absoluta y donde existe capacidad técnica para implementar, ajustar y mantener el modelo. Es particularmente adecuado en instituciones educativas de gran tamaño que ya cuentan con equipos de analítica de datos y que buscan optimizar la detección temprana de estudiantes en riesgo, incluso a costa de sacrificar cierta interpretabilidad (Harnisher et al., 2024). En escenarios de investigación, XGBoost es útil para establecer cotas superiores de rendimiento predictivo contra las cuales comparar modelos más interpretables. No obstante, su aplicabilidad directa en servicios de orientación vocacional de menor escala o con recursos técnicos limitados es cuestionable, ya que la complejidad del modelo puede exceder la capacidad de implementación y mantenimiento de la institución.

Redes Neuronales

Las redes neuronales artificiales (RNA), particularmente en sus variantes profundas de deep learning, ofrecen una capacidad teórica inigualable para modelar relaciones extremadamente complejas y no lineales, así como para aprender representaciones jerárquicas de los datos sin necesidad de ingeniería de variables explícita (Haykin, 2009). En dominios como el reconocimiento de imágenes o el procesamiento del lenguaje natural, las redes neuronales han logrado avances revolucionarios. En principio, esta capacidad podría ser relevante para la predicción de deserción si se dispusiera de datos educativos masivos, heterogéneos y

longitudinales tales como registros de interacciones en plataformas digitales, textos de foros, patrones de navegación.(Tsai et al., 2020)

En la revisión las redes neuronales están presentes en el **14%** (5/35) de los estudios, sin embargo, en general, las redes neuronales han mostrado resultados modestos y colectivamente inferiores a Random Forest y XGBoost en los estudios revisados. (Dake & Buabeng-Andoh, 2022) reportaron que un perceptrón multicapa (MLP) alcanzó una precisión del 68.2% y un F1-score de 0.64, en ambos casos por debajo de Random Forest (70.98% y 0.69). Las razones son múltiples. Primero, las redes neuronales requieren volúmenes de datos del orden de decenas de miles de registros para entrenarse adecuadamente, mientras que la mayoría de los estudios educativos trabajan con muestras de pocos miles o incluso cientos de estudiantes. Segundo, son extremadamente sensibles a la escala de las variables y requieren una normalización o estandarización cuidadosa. Tercero, su naturaleza de "caja negra" es aún más opaca que la de XGBoost, lo que dificulta enormemente la interpretación y la comunicación de resultados a orientadores y gestores (Rodriguez Ramirez et al., 2022). Cuarto, su entrenamiento es computacionalmente costoso y requiere hardware especializado (GPUs) para ser práctico con conjuntos de datos grandes.

En el estado actual de la literatura, las redes neuronales no se recomiendan mucho para la predicción de deserción universitaria en la mayoría de los contextos institucionales, especialmente cuando se dispone de datos de menos de 10.000 registros y cuando la interpretabilidad es una preocupación. Su aplicabilidad podría reconsiderarse en el futuro, a medida que las instituciones educativas acumulen grandes volúmenes de datos como registros de interacción minuto a minuto en plataformas de aprendizaje y que las técnicas de IA explicable (XAI) permitan abrir la "caja negra" de las redes neuronales. Por ahora, los métodos de ensemble

como Random Forest y XGBoost ofrecen un mejor equilibrio entre rendimiento, interpretabilidad y facilidad de implementación (Eegdeman, 2023).

La siguiente tabla sintetiza la comparación multidimensional de las cinco técnicas analizadas, permitiendo una visualización rápida de sus perfiles diferenciales.

Tabla 5

Matriz Comparativa de Modelos Predictivos

Técnica	Ventajas principales	Limitaciones principales	Aplicabilidad recomendada
Random Forest	Alto rendimiento, robustez al sobreajuste, manejo de alta dimensionalidad, importancia de variables	Interpretabilidad media (caja gris), costo computacional moderado, requiere ajuste de hiperparámetros	Instituciones medianas/grandes con personal técnico; prioridad en precisión predictiva; educación a distancia
Regresión Logística	Alta interpretabilidad (odds ratios), computacionalmente liviana, fácil implementación, buen baseline	Supuesto de linealidad, sensible a multicolinealidad y atípicos, rendimiento inferior en relaciones no lineales	Contextos con recursos técnicos limitados; prioridad en comprensión de factores; comunicación a audiencias no técnicas
Árboles de Decisión	Transparencia gráfica, intuitivos, sin supuestos distribucionales, robustos a datos faltantes	Tendencia al sobreajuste, inestabilidad ante variaciones en datos, rendimiento inferior a ensemble	Entornos donde la comprensión del proceso es prioritaria; apoyo a entrevistas de orientación; instituciones con recursos

Técnica	Ventajas principales	Limitaciones principales	Aplicabilidad recomendada
			básicos
XGBoost	Máximo rendimiento predictivo, manejo eficiente de desbalance, importancia de variables	Baja interpretabilidad (caja negra), alta complejidad técnica, curva de aprendizaje pronunciada, costo computacional alto	Instituciones grandes con equipos de analítica; máxima precisión como prioridad; investigación de frontera
Redes Neuronales	Capacidad teórica para relaciones muy complejas, aprendizaje de representaciones	Rendimiento modesto en la práctica, requiere grandes volúmenes de datos, caja negra, costo computacional muy alto	No recomendado en el estado actual; potencial futuro con datos masivos y XAI

Nota. La tabla detalla una comparativa de ventajas, limitaciones y aplicabilidad de modelos predictivos en contextos educativos en aprendizaje automático aplicado a educación.

Resultado 5 Lineamientos Teóricos para el Uso de Modelos Predictivos en Decisiones Vocacionales

Se formulan a continuación un conjunto de lineamientos orientados a guiar la implementación de modelos predictivos como herramientas de apoyo en los procesos de orientación vocacional. Estos lineamientos no pretenden constituir un manual técnico de implementación, sino más bien un marco conceptual y estratégico que dialogue con la literatura especializada y que pueda ser adaptado a diferentes contextos institucionales. Se estructuran en cinco dimensiones interrelacionadas: (i) fundamentos conceptuales y éticos, (ii) selección de

técnicas y métricas, (iii) integración de variables vocacionales, (iv) diseño de intervenciones basadas en predicciones, y (v) validación y mejora continua.

Fundamentos Conceptuales y Éticos

Principio 1. Los modelos predictivos son herramientas de apoyo, no sustitutos del juicio profesional**.** La literatura revisada es unánime en señalar que los modelos predictivos no deben reemplazar la labor de los orientadores vocacionales, sino complementarla y potenciarla (Eegdeman, 2023; Rodriguez Ramirez et al., 2022). Un modelo puede identificar a un estudiante con alta probabilidad de deserción basándose en patrones históricos, pero no puede sustituir la comprensión contextual, la empatía y el juicio clínico que un orientador aporta en una entrevista personalizada. Por lo tanto, se propone que la salida de los modelos se conciba como una señal de alerta o insumo para la priorización, no como un diagnóstico definitivo. El orientador debe tener siempre la facultad de sobreseer la predicción del modelo cuando su evaluación profesional indique que el contexto del estudiante justifica una decisión diferente.

Principio 2. Los modelos deben diseñarse y utilizarse con salvaguardas éticas explícitas. La aplicación de algoritmos predictivos en educación conlleva riesgos éticos significativos que deben ser abordados desde el diseño mismo del sistema (Harnisher et al., 2024) En primer lugar, el sesgo algorítmico puede reproducir o incluso amplificar desigualdades preexistentes si los datos de entrenamiento reflejan discriminaciones históricas, por ejemplo, si ciertos grupos socioeconómicos o étnicos han tenido sistemáticamente menos acceso a recursos de apoyo. Se recomienda auditar periódicamente los modelos para detectar desempeños diferenciales entre subgrupos poblacionales. En segundo lugar, la privacidad y confidencialidad de los datos estudiantiles debe garantizarse mediante anonimización, control de accesos y consentimiento informado. Los estudiantes y sus familias deben conocer qué datos se recogen, cómo se utilizan

y con qué fines. En tercer lugar, debe evitarse la estigmatización de los estudiantes señalados como "en riesgo"; las predicciones deben manejarse con cuidado comunicacional y utilizarse exclusivamente para ofrecer apoyo, nunca para sancionar o excluir.

Principio 3. La orientación vocacional debe mantenerse como marco de referencia. Los modelos predictivos, por sí mismos, no resuelven el problema de la deserción. Su valor se realiza plenamente cuando se insertan en un sistema de orientación vocacional que incluye acompañamiento personalizado previo al ingreso, evaluación de intereses, aptitudes y autoeficacia, seguimiento durante los primeros semestres, e intervenciones diferenciadas según el perfil de riesgo. Los modelos predictivos pueden informar cada una de estas etapas, pero no reemplazan la necesidad de una estructura institucional de apoyo (Lent et al., 1994; Savickas, 2019).

Selección de Técnicas y Métricas Según el Propósito y el Contexto Institucional

Principio 1. La elección del modelo debe equilibrar rendimiento predictivo e interpretabilidad. A partir de la comparación realizada en el Objetivo Específico 4, se propone la siguiente matriz de decisión para la selección de técnicas:

Tabla 6

Matriz de Decisión para la Selección de Técnicas

Propósito principal	Contexto institucional	Técnica recomendada	Justificación
Maximizar precisión predictiva	Institución grande, con equipo técnico	Random Forest o XGBoost	Mayor AUC-ROC y F1-score; capacidad técnica disponible
Comprender factores de riesgo	Recursos técnicos limitados; prioridad	Regresión Logística	Odds ratios comprensibles para audiencias no técnicas

Propósito principal	Contexto institucional	Técnica recomendada	Justificación
	interpretativa		
Apoyar entrevistas de orientación	Servicio de orientación con pocos datos	Árboles de Decisión	Transparencia gráfica; puede recorrerse visualmente con el estudiante
Investigación / benchmarking	Cualquier contexto con suficientes datos	Múltiples modelos (baseline + ensemble)	Permite comparación y establecimiento de cotas superiores

Nota. La tabla muestra matriz de decisión para la selección de técnicas según propósito y contexto institucional.

Principio 2. Las métricas de evaluación deben seleccionarse según el costo relativo de los errores. Dado que el costo de un falso negativo (no detectar a un desertor real) es generalmente mayor que el de un falso positivo (señalar erróneamente a un estudiante como en riesgo), se recomienda priorizar métricas sensibles al desempeño en la clase minoritaria (Eegdeman, 2023). Específicamente, se propone: F1-score como métrica principal de reporte, por su capacidad de equilibrar precisión y sensibilidad. AUC-ROC como métrica secundaria, por su independencia del umbral de clasificación. Sensibilidad (recall) como métrica de seguimiento, estableciendo un umbral mínimo aceptable (ej. ≥ 0.70) para considerar que el modelo es útil operativamente.Exactitud (accuracy) solo como métrica complementaria, con la advertencia explícita de su posible carácter engañoso en contextos con desbalance de clases.

Principio 3. El tratamiento del desbalance de clases es obligatorio. Dado que la deserción suele afectar a una minoría de estudiantes (entre el 10% y el 40% según los estudios revisados), se recomienda aplicar al menos una de las siguientes estrategias antes del entrenamiento del modelo: sobremuestreo de la clase minoritaria mediante SMOTE (Synthetic Minority Over-sampling Technique), submuestreo de la clase mayoritaria, o uso de pesos de clase en el algoritmo (James et al., 2021) La efectividad de estas técnicas debe evaluarse comparando el F1-score y la sensibilidad con y sin su aplicación.

Integración de Variables Vocacionales en los Modelos Predictivos

Principio 1. Los modelos predictivos de deserción deben incorporar variables psicológicas y vocacionales, no solo académicas y socioeconómicas. Una limitación recurrente en el estado del arte, identificada, es la escasa inclusión de variables vocacionales en los modelos predictivos. La evidencia revisada (Rodriguez Ramirez et al., 2022; Zhu et al., 2019) demuestra que la incorporación de variables como autoeficacia vocacional, indecisión profesional, satisfacción con la carrera elegida y adaptabilidad profesional (career adaptability) mejora significativamente el rendimiento predictivo, elevando el AUC-ROC en 0.05 a 0.10 puntos en comparación con modelos que solo usan variables demográficas y académicas.

Principio 2. Se propone un conjunto mínimo de variables vocacionales para recoger sistemáticamente. Con base en la literatura de orientación vocacional (Cunningham et al., 1977; Lent et al., 1994; Savickas, 2019) y en los hallazgos revisados, se sugiere que las instituciones educativas que aspiren a implementar modelos predictivos de deserción incluyan, al menos, los siguientes constructos en sus sistemas de recolección de datos:

Tabla 7

Constructos para Recolección de Datos

Constructo	Instrumento sugerido	Fuente teórica	Relación con deserción
Autoeficacia vocacional	Escala de Autoeficacia para Elección de Carrera (adaptada de (Lent et al., 1994).)	Lent, Brown & Hackett (1994)	A menor autoeficacia, mayor probabilidad de deserción
Indecisión vocacional	Career Decision Scale (CDS) o versión breve	(Cunningham et al., 1977)	Alta indecisión se asocia con cambio de carrera o abandono
Satisfacción con la carrera elegida	Ítem único o escala breve (ej. "Estoy satisfecho/a con mi elección de carrera")	Adaptado de (Rodriguez Ramirez et al., 2022)Haga clic o pulse aquí para escribir texto.	Baja satisfacción predictora de deserción temprana
Adaptabilidad profesional (career adaptability)	Career Adapt-Abilities Scale (CAAS) versión corta	(Savickas, 2019)	Mayor adaptabilidad se asocia con menor intención de abandono (Zhu et al., 2019)
Congruencia persona-carrera	Cálculo basado en código Holland del estudiante y código de la carrera	(Cunningham et al., 1977)	Baja congruencia predictora de insatisfacción y deserción

Nota. La tabla detalla los constructos recomendados para recolección de datos.

Principio 3. Las variables vocacionales deben recogerse en dos momentos temporales. Dado que la autoeficacia, la satisfacción y la adaptabilidad pueden cambiar durante la experiencia universitaria (Lent et al., 1994), se recomienda administrar los instrumentos en al menos dos momentos: al ingreso (línea base vocacional) y al final del primer semestre (evaluación del desajuste temprano). La comparación entre ambos momentos puede generar variables de cambio que son especialmente predictivas de deserción (Eegdeman, 2023).

Diseño de Intervenciones Basadas en Predicciones

Principio 1. Las predicciones deben traducirse en niveles de riesgo con umbrales claros. En lugar de entregar a los orientadores una probabilidad continua de deserción (ej. "0.73"), se recomienda establecer categorías de riesgo que orienten la toma de decisiones. Una propuesta viable, basada en la práctica reportada por (Harnisher et al., 2024), es la siguiente:

Tabla 8

Propuesta de Categorías de Riesgos

Probabilidad	Riesgo	Intervención sugerida
< 0.30	Bajo	Seguimiento estándar; sin intervención específica
0.30 - 0.60	Moderado	Monitoreo; entrevista breve de verificación; oferta de talleres de apoyo
0.60 - 0.80	Alto	Entrevista de orientación en profundidad; plan de acompañamiento personalizado
> 0.80	Muy alto	Intervención inmediata; derivación a servicios de consejería; revisión de adecuación carrera-estudiante

Nota. La tabla muestra las categorías de riesgos, los umbrales deben calibrarse empíricamente según las características de cada institución y las tasas base de deserción.

Principio 2. Las intervenciones deben ser diferenciadas según el perfil de riesgo identificado por el modelo. Un modelo predictivo que incluye variables vocacionales permite no solo estimar la probabilidad de deserción, sino también identificar qué factores específicos están impulsando ese riesgo. Por ejemplo, un estudiante con alta probabilidad de deserción impulsada principalmente por baja autoeficacia vocacional requerirá una intervención distinta (talleres de desarrollo de autoeficacia, mentorías con pares exitosos) que un estudiante cuyo riesgo se debe fundamentalmente a baja congruencia persona-carrera (reorientación vocacional, exploración de alternativas curriculares). La especificidad de la intervención, basada en las variables más importantes para cada caso, es una de las principales ventajas de integrar modelos predictivos con fundamentos de orientación vocacional (Rodriguez Ramirez et al., 2022).

Principio 3. La comunicación de las predicciones a los estudiantes debe ser cuidadosa y ética. Los estudiantes tienen derecho a conocer si han sido identificados como "en riesgo", pero esta comunicación debe manejarse con sensibilidad para evitar efectos contraproducentes (estigmatización, profecía autocumplida). Se recomienda que la comunicación sea realizada por un orientador entrenado, en el contexto de una relación de confianza, y que se enmarque como una oportunidad de apoyo, no como una etiqueta o un diagnóstico negativo. El mensaje sugerido podría ser: "Según nuestros análisis, algunos estudiantes con un perfil similar al tuyo han enfrentado dificultades para continuar sus estudios. Queremos ofrecerte recursos adicionales para que puedas tener éxito" (Harnisher et al., 2024).

Validación y Mejora Continua de los Modelos

Principio 1. Los modelos deben validarse externamente antes de su implementación operativa. Una limitación recurrente en la literatura es la falta de validación externa de los modelos propuestos (Eegdeman, 2023). Un modelo entrenado con datos de una cohorte de

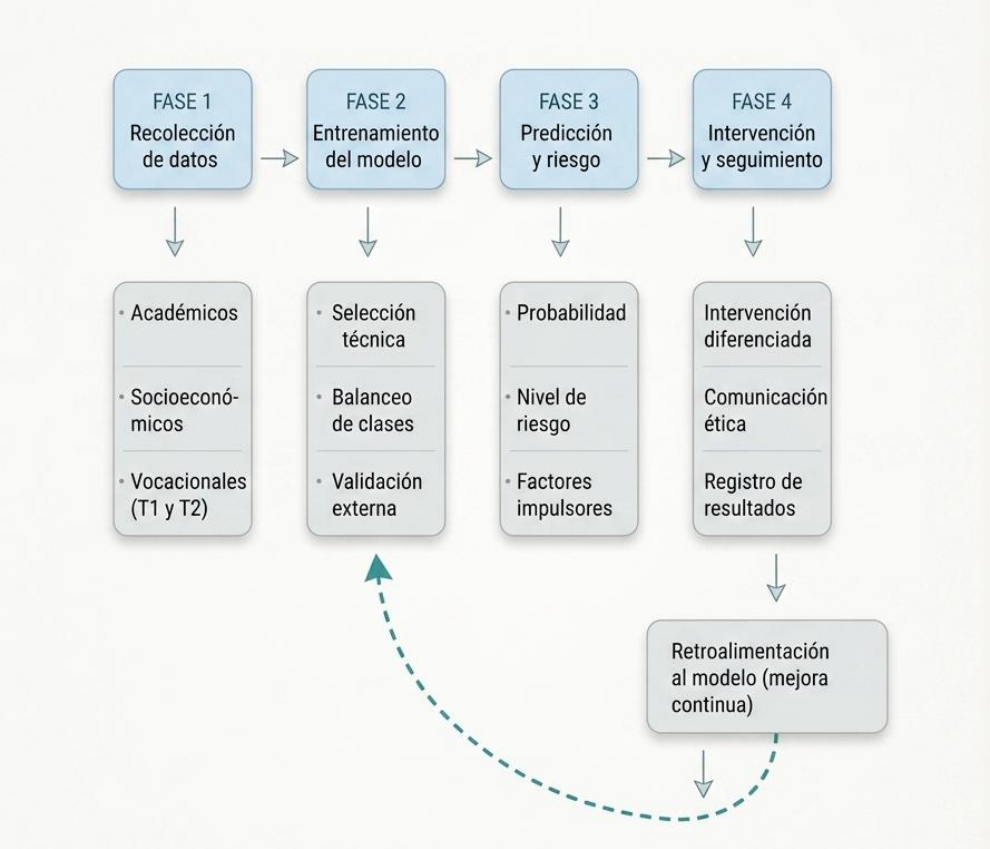
estudiantes puede no generalizar bien a cohortes posteriores o a otras instituciones. Se recomienda que, antes de implementar un modelo de manera operativa, se realice una validación temporal (entrenar con datos de cohortes pasadas y probar con la cohorte más reciente) y, si es posible, una validación institucional cruzada. Solo cuando el modelo demuestre estabilidad en estas validaciones externas se justifica su uso rutinario.

Principio 2. Los modelos deben ser recalibrados periódicamente. Las dinámicas de deserción universitaria no son estáticas; cambian con las políticas institucionales, las características de las cohortes entrantes, las condiciones socioeconómicas y hasta eventos exógenos como una pandemia. Por lo tanto, se recomienda establecer un cronograma de recalibración donde el modelo se reentrene con los datos más recientes. También se sugiere implementar un sistema de monitoreo de deriva del modelo (model drift detection) que alerte cuando el desempeño del modelo caiga por debajo de umbrales preestablecidos (James et al., 2021)

Principio 3. Los resultados de las intervenciones deben retroalimentar el modelo. Un sistema de mejora continua requiere que la información sobre qué intervenciones se realizaron y con qué resultados se incorpore al modelo. Por ejemplo, si un estudiante identificado como "en riesgo" recibe una intervención de tutoría y logra permanecer, ese dato debe alimentar futuras iteraciones del modelo, permitiendo que aprenda qué combinaciones de perfil de riesgo e intervención son más efectivas. Este enfoque, conocido como aprendizaje activo o sistemas de recomendación adaptativos, representa una dirección prometedora para la integración de modelos predictivos y orientación vocacional (Eegdeman, 2023; Rodriguez Ramirez et al., 2022).

Figura 1

Modelo Integrador



Nota. En la imagen se aprecia modelo integrador para el uso de modelos predictivos en orientación vocacional.

Discusión

La revisión sistemática de 35 estudios confirma que los modelos predictivos basados en aprendizaje supervisado ofrecen un potencial real para apoyar la orientación vocacional y reducir la deserción universitaria, pero su efectividad depende de equilibrar rendimiento e interpretabilidad, incorporar variables psicoeducativas y adaptar los modelos al contexto institucional.

Equilibrio entre Precisión e Interpretabilidad

Los resultados muestran que no existe un modelo universalmente superior. La regresión logística (29% de los estudios) y Random Forest (26%) son las técnicas más frecuentes, pero con perfiles opuestos. Random Forest alcanza precisiones del 70-85% y es robusto frente al sobreajuste, aunque su interpretabilidad es media ("caja gris"). La regresión logística, con precisiones del 65-75%, permite obtener *odds ratios* fácilmente comprensibles para orientadores y gestores (Hosmer et al., 2013). Los árboles de decisión (17%) ofrecen un punto intermedio mediante la representación gráfica y capacidad predictiva aceptable. Las redes neuronales (14%) no mostraron ventajas consistentes, probablemente porque los conjuntos de datos educativos típicos aún son modestos para arquitecturas profundas (Tsai et al., 2020). Por tanto, las instituciones deben elegir el modelo según su prioridad. Si se busca máxima detección temprana de riesgo, Random Forest es la mejor opción. Si se necesita comprender los factores para diseñar intervenciones personalizadas, la regresión logística o los árboles de decisión resultan más adecuados.

Infrarrepresentación de Variables Vocacionales

Solo el 23% de los estudios (8 de 35) incorporan datos psicoeducativos como autoeficacia, indecisión vocacional, satisfacción con la carrera o adaptabilidad profesional. La

mayoría se limita a variables académicas y socioeconómicas. Esto es problemático porque, como demuestran Zhu et al. (2019) y Coetzee et al. (2023), la adaptabilidad profesional y la autoeficacia tienen un efecto negativo directo sobre la intención de abandono. Rodriguez Ramirez et al. (2022) mostraron que añadir datos de asesoría estudiantil elevaba el AUC-ROC de 0,78 a 0,85. En este orden de ideas, las instituciones deben recolectar estas variables al menos en dos momentos (ingreso y final del primer semestre) para alimentar modelos predictivos y generar alertas tempranas. Para la UNAD, se recomienda comenzar con árboles de decisión que incluyan variables de interacción en el campus virtual (accesos, entregas, participación en foros), tal como hizo Avila Pérez (2021).

Limitaciones del Estudio

Deben reconocerse varias limitaciones, una es que más del 70% de los estudios provienen de contextos presenciales en países desarrollados, lo que limita la generalización a la educación a distancia latinoamericana; otra es que las definiciones de deserción varían (abandono en primer año, cambio de carrera, deserción institucional), dificultando comparaciones cuantitativas; otra limitación es que solo seis estudios reportaron validación externa, lo que indica que la mayoría de los modelos no han sido probados en poblaciones diferentes a las de entrenamiento.

Futuras Líneas de Investigación

Se requieren estudios cuasiexperimentales que comparen la efectividad de modelos predictivos con y sin variables vocacionales en la reducción real de la deserción. También es necesaria investigación aplicada en entornos de educación a distancia latinoamericanos o en Colombia, donde las plataformas como la de la UNAD generan abundantes datos de interacción aún no aprovechados. Por último, el avance de la inteligencia artificial explicable (XAI) podría permitir el uso de redes neuronales profundas sin sacrificar la interpretabilidad.

Conexión con los Lineamientos Propuestos

La discusión valida los cinco lineamientos presentados: integrar variables vocacionales, usar métricas robustas (F1-score, AUC-ROC), tratar el desbalance de clases (SMOTE), diseñar intervenciones según niveles de riesgo y recalibrar los modelos periódicamente. Constituyen una hoja de ruta válidas para que instituciones como la UNAD implementen sistemas de apoyo a la permanencia basados en datos, sin reemplazar el juicio profesional de los orientadores.

Conclusiones

El presente trabajo se propuso analizar los modelos predictivos basados en técnicas de aprendizaje supervisado aplicados a la orientación vocacional, con el fin de identificar su pertinencia, ventajas y limitaciones como apoyo en la elección de carrera universitaria y en la reducción de la deserción. Este ejercicio permitió construir una visión integral del estado del arte, las técnicas disponibles, las métricas de evaluación y los lineamientos para una implementación adecuada.

Se concluye que existe un creciente interés académico en la aplicación de técnicas de aprendizaje automático al ámbito educativo, con un notable incremento de publicaciones en el período 2010 -2026. Sin embargo, la revisión de los estudios incluidos revela una brecha en cuanto a que el 69% de las investigaciones se centran exclusivamente en la predicción de la deserción universitaria a partir de variables académicas y socioeconómicas, mientras que solo el 11% abordan explícitamente la orientación vocacional o la recomendación de carreras, y otro 20% tratan fenómenos relacionados (intención de abandono, adaptabilidad profesional) que pueden vincularse con la orientación. Lo anterior valida la pertinencia del presente trabajo y evidencia una oportunidad de investigación desatendida.

La regresión logística (29%) y Random Forest (26%) son las técnicas más frecuentes, siendo Random Forest la que presenta el mejor desempeño predictivo reportado en las comparativas. La regresión logística y los árboles de decisión (17%) mantienen un lugar relevante debido a su alta interpretabilidad, lo que los hace especialmente valiosos en contextos donde la comprensión de los factores de riesgo es tan importante como la predicción misma. XGBoost (11%) representa la frontera del rendimiento predictivo, aunque con mayores requerimientos técnicos. Las redes neuronales (14%), pese a su sofisticación, no han demostrado

ventajas consistentes en este dominio, probablemente debido a los tamaños muestrales moderados que caracterizan a la mayoría de los estudios educativos.

La evaluación rigurosa de modelos predictivos en contextos de deserción universitaria no puede reducirse a una única métrica. La exactitud (accuracy), aunque es la más reportada (86% de los estudios), puede ser engañosa en contextos con desbalance de clases. Por el contrario, el F1-score (reportado en el 63% de los estudios) y el AUC-ROC (69%) emergen como las métricas más robustas, pues evalúan el equilibrio entre precisión y sensibilidad y la capacidad discriminativa del modelo independientemente del umbral de clasificación, respectivamente. Se concluye, además, que el tratamiento del desbalance de clases mediante técnicas como SMOTE es una práctica recomendada que mejora la sensibilidad y el F1-score.

También se concluye que no existe un modelo universalmente superior. La elección de la técnica más adecuada depende de un balance entre múltiples factores: el rendimiento predictivo requerido, la necesidad de interpretabilidad, los recursos técnicos disponibles, el tamaño y la calidad de los datos, y el propósito último del modelo. Random Forest ofrece el mejor equilibrio para la mayoría de los contextos institucionales, combinando un alto rendimiento predictivo con una interpretabilidad aceptable y requerimientos técnicos moderados. La regresión logística y los árboles de decisión son preferibles cuando la transparencia y la comunicación a audiencias no técnicas constituyen prioridades centrales. XGBoost se recomienda para instituciones con equipos de analítica consolidados que buscan maximizar la precisión predictiva. Las redes neuronales, por ahora, no se recomiendan para aplicaciones prácticas en este dominio.

la integración de modelos predictivos con los fundamentos de la orientación vocacional no solo es posible, sino una necesidad. Los cinco lineamientos propuestos: fundamentos conceptuales y éticos, selección de técnicas y métricas, integración de variables vocacionales,

diseño de intervenciones basadas en predicciones y validación y mejora continua, constituyen un marco integrador que trasciende el mero análisis técnico para ofrecer una guía aplicable a instituciones educativas de diverso tamaño y recursos. Especial relevancia tiene la recomendación de incorporar variables vocacionales como autoeficacia, indecisión, satisfacción con la carrera, adaptabilidad profesional y congruencia persona-carrera, cuyo valor predictivo ha sido demostrado en la literatura revisada pero cuya inclusión sistemática sigue siendo poca en la práctica.

Finalmente, los modelos predictivos, particularmente aquellos basados en aprendizaje supervisado, ofrecen un potencial significativo para apoyar los procesos de orientación vocacional y contribuir a la reducción de la deserción universitaria. Sin embargo, su valor no reside en la predicción misma, sino en su capacidad para generar información oportuna que oriente intervenciones diferenciadas, personalizadas y preventivas. La integración de variables vocacionales en los modelos, el uso de métricas robustas como el F1-score y el AUC-ROC, la selección de técnicas que equilibren rendimiento e interpretabilidad según el contexto, y el diseño de intervenciones éticamente fundamentadas son condiciones necesarias para que este potencial se traduzca en beneficios reales para los estudiantes. Este estudio pretende ser la puerta de entrada para futuras investigaciones y desarrollos aplicados en el ámbito de la orientación vocacional predictiva.

Recomendaciones

A partir de las conclusiones derivadas del presente trabajo, y con el propósito de contribuir a la mejora de las prácticas institucionales y al avance del conocimiento en el campo, se formulan las siguientes recomendaciones dirigidas a instituciones de educación superior, orientadores vocacionales y servicios de apoyo estudiantil, y futuros investigadores.

Se recomienda que las instituciones incorporen en sus procesos de admisión y seguimiento la recolección estructurada de datos vocacionales: autoeficacia, indecisión, satisfacción con la carrera elegida, adaptabilidad profesional y congruencia persona-carrera, utilizando instrumentos validados como la Career Decision Scale, la Career Adapt-Abilities Scale (CAAS) o adaptaciones de las escalas de Lent, Brown y Hackett. Estos datos deben recogerse al menos en dos momentos: al ingreso y al final del primer semestre.

Se recomienda que las instituciones, particularmente aquellas de tamaño mediano y grande, inviertan en la formación de personal técnico (analistas de datos, científicos de datos) o en alianzas con unidades académicas afines, que permitan implementar y mantener modelos predictivos. Random Forest se propone como punto de partida recomendado por su equilibrio entre rendimiento e interpretabilidad.

Antes de implementar cualquier sistema predictivo de manera operativa, las instituciones deben desarrollar protocolos explícitos que aborden la auditoría periódica de sesgos algorítmicos, la garantía de privacidad y confidencialidad de los datos estudiantiles, el consentimiento informado de los estudiantes, y la prohibición explícita de utilizar las predicciones para excluir estudiantes. Estos protocolos deben ser revisados por comités de ética institucionales.

La calidad de los modelos predictivos depende críticamente de la calidad y completitud de los datos. Se recomienda que las instituciones fortalezcan sus sistemas de información para

registrar de manera sistemática trayectorias académicas, interacciones con plataformas virtuales, participación en actividades de apoyo y, crucialmente, las intervenciones realizadas y sus resultados, para permitir la retroalimentación y mejora continua de los modelos.

Los orientadores deben concebir las predicciones como señales de alerta o insumos para la priorización de casos, no como diagnósticos definitivos. La decisión final sobre qué intervención realizar, o incluso si se debe intervenir, debe permanecer en el ámbito del juicio profesional del orientador, quien considera factores contextuales que el modelo no puede capturar.

Dada su transparencia gráfica, los árboles de decisión pueden utilizarse durante las entrevistas con estudiantes para explorar visualmente las condiciones que aumentan o disminuyen el riesgo de deserción, facilitando la identificación conjunta de áreas de intervención. Esta herramienta es especialmente útil en contextos con recursos técnicos limitados.

La principal brecha identificada en este estudio es la escasa incorporación de datos de orientación vocacional en los modelos predictivos de deserción toda vez que solo 8 de los 35 estudios (23%) incluyen variables psicoeducativas como autoeficacia, indecisión o adaptabilidad profesional.

Referencias Bibliográficas

Abadía, C., Guerrero, H., Rodríguez, J., Vela González, P. A., Martínez, H., Villamizar, A. N., Sánchez, D. A. A., & Aguilar, L. A. P. (2018). *Informes de Gestión de Procesos -IGP- vigencia 2018 - Ciclo Vida Del Estudiante*. https://sig.unad.edu.co/documentos/sgc/informes_gestion/2018/periodo_1/1er_IGP_2018_ciclo_de_vida_del_estudiante.pdf

Avila Pérez, M. L. (2021). *Modelo de predicción de deserción estudiantil, apoyado en Tecnologías De Data Mining, en un curso de primera matrícula de la Escuela ECBTI De La UNAD.* http://repository.unad.edu.co/handle/10596/42544

Avila Pérez, M. L., & Medina, J. (2020). Minería de datos para la predicción de la deserción estudiantil en la Universidad Nacional Abierta y a Distancia. *Documentos de Trabajo ECBTI*, *1*(2). https://doi.org/10.22490/ECBTI.4354

Berens, J., Schneider, K., Gortz, S., Oster, S., & Burghoff, J. (2019). Early Detection of Students at Risk -- Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods. *Journal of Educational Data Mining*, *11*(3), 1–41. https://doi.org/https://doi.org/10.5281/zenodo.3594771

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

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). Classification and regression trees. In *Classification and Regression Trees*. CRC Press. https://doi.org/10.1201/9781315139470

Cabus, S. J., & De Witte, K. (2016). Why Do Students Leave Education Early? Theory and Evidence on High School Dropout Rates. *Journal of Forecasting*, *35*(8), 690–702. https://doi.org/10.1002/FOR.2394

Cedillo - Quizhpe, C., Arias Blanco, J. M., & Burguera Condon, J. L. (2026). Influencia de las expectativas académicas sobre los motivos de intención de abandono en estudiantes de una universidad ecuatoriana. *Educación XX1*, *29*(1), 95–122. https://doi.org/10.5944/EDUCXX1.42831

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 785–794. https://doi.org/10.1145/2939672.2939785;CSUBTYPE:STRING:CONFERENCE

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics 2019 21:1*, *21*(1), 6-. https://doi.org/10.1186/S12864-019-6413-7

Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F., & Peter Campbell, J. (2020). Introduction to machine learning, neural networks, and deep learning. *Translational Vision Science and Technology*, *9*(2). https://doi.org/10.1167/tvst.9.2.14

Coetzee, M., Mbiko, H. N., & Nel, E. (2023). To what extent do career agility and psychological capital activate employees' career adaptability and foster their career resilience and career satisfaction? *South African Journal of Psychology*, *53*(3). https://doi.org/10.1177/00812463231186271

Cunningham, C. H., Alston, H. L., Doughtie, E. B., & Wakefield, J. A. (1977). Use of Holland's vocational theory with potential high school dropouts. *Journal of Vocational Behavior*, *10*(1), 35–38. https://doi.org/10.1016/0001-8791(77)90039-2

Dake, D. K., & Buabeng-Andoh, C. (2022). Using Machine Learning Techniques to Predict Learner Drop-out Rate in Higher Educational Institutions. *Mobile Information Systems*, *2022*(1), 2670562. https://doi.org/10.1155/2022/2670562

Eegdeman, I. M. (2023). *Enhancing Study Success in Dutch Vocational Education* [Vrije Universiteit Amsterdam]. https://doi.org/10.5463/THESIS.100

Facundo Diaz, A. H. (2009). Análisis sobre la deserción en la educación superior a distancia y virtual: el caso de la UNAD - COLOMBIA. *Revista de Investigaciones UNAD*, *8*(2), 117–149. https://doi.org/10.22490/25391887.639

Federici, E., Boon, C., & Den Hartog, D. N. (2021). The moderating role of HR practices on the career adaptability–job crafting relationship: a study among employee–manager dyads. *International Journal of Human Resource Management*, *32*(6). https://doi.org/10.1080/09585192.2018.1522656

García-Botero, L., Álvarez-Maestre, A. J., Pérez-Fuentes, C. A., Rodriguez, C.-M. C., & Johana Aguilar-Barreto, A. (2022). REVISIÓN SISTEMÁTICA: CRITERIOS DE CALIDAD EN EL PROYECTO DE PROGRAMAS DE ORIENTACIÓN VOCACIONAL. *Psicologia Escolar e Educacional*, *26*. https://doi.org/10.1590/2175-35392022-235549

Ghorbani, R., & Ghousi, R. (2020). Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques. *IEEE Access*, *8*, 67899–67911. https://doi.org/10.1109/ACCESS.2020.2986809

Gochhayat, N., & Ravindran, R. (2025). Drop out v/s retention of female students: unfolding dynamics of the education system in Indian states. *Discover Education 2025 4:1*, *4*(1), 176-. https://doi.org/10.1007/S44217-025-00552-0

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques*. Morgan Kaufmann. http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

Harnisher, J., Villanueva, S., Prieto, D., Anzaldua, R., Beem, K., Harnisher, J., Villanueva, S., Prieto, D., Anzaldua, R., & Beem, K. (2024). Collaborative Development of Machine Learning Algorithms for Student Success at John Jay College. *Journal of Access, Retention, and Inclusion in Higher Education*, *7*(1), 2. https://digitalcommons.wcupa.edu/jarihe/vol7/iss1/2

Haykin, S. (2009). *Neural Networks and Learning Machines. Third Edition, Pearson Education, Inc.* (C. McMaster University, Ed.; 3rd edition). Pearson. https://dai.fmph.uniba.sk/courses/NN/haykin.neural-networks.3ed.2009.pdf

Heredia, D., Amaya, Y., & Barrientos, E. (2015). Student Dropout Predictive Model Using Data Mining Techniques. *IEEE Latin America Transactions*, *13*(9), 3127–3134. https://doi.org/10.1109/TLA.2015.7350068

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression: Third Edition. *Applied Logistic Regression: Third Edition*, 1–510. https://doi.org/10.1002/9781118548387

Huamán Arratia, H. H. (2025). Modelo predictivo para la prevención de la deserción académica en estudiantes de un instituto tecnológico de Puno, 2025. *Journal of Scientific and Technological Research Industrial*, *6*(2), 35–58. https://doi.org/10.47422/jstri.v6i2.67

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning : with Applications in R* (D. of S. G. Allen, Ed.; Segunda Edición). Springer US. https://doi.org/10.1007/978-1-0716-1418-1

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An Introduction to Statistical Learning: With Applicationn in Python*. Springer International Publishing. https://doi.org/10.1007/978-3-031-38747-0

Johnson, J. A. (2014). The Ethics of Big Data in Higher Education. *The International Review of Information Ethics*, *21*(07), 3–10. https://doi.org/10.29173/IRIE365

Kabathova, J., & Drlik, M. (2021). Towards Predicting Student's Dropout in University Courses Using Different Machine Learning Techniques. *Applied Sciences 2021, Vol. 11, Page 3130*, *11*(7), 3130. https://doi.org/10.3390/APP11073130

Kemper, L., Vorhoff, G., & Wigger, B. U. (2020). Predicting student dropout: A machine learning approach. *European Journal of Higher Education*, *10*(1), 28–47. https://doi.org/10.1080/21568235.2020.1718520

Kocsis, Á., & Molnár, G. (2025). Factors influencing academic performance and dropout rates in higher education. *Oxford Review of Education*, *51*(3), 414–432. https://doi.org/10.1080/03054985.2024.2316616

Lent, R. W., Brown, S. D., & Hackett, G. (1994). Toward a Unifying Social Cognitive Theory of Career and Academic Interest, Choice, and Performance. *Journal of Vocational Behavior*, *45*(1), 79–122. https://doi.org/10.1006/JVBE.1994.1027

Macarini, L. A. B., Cechinel, C., Machado, M. F. B., Ramos, V. F. C., & Munoz, R. (2019). Predicting Students Success in Blended Learning—Evaluating Different Interactions Inside Learning Management Systems. *Applied Sciences 2019, Vol. 9, Page 5523*, *9*(24), 5523. https://doi.org/10.3390/APP9245523

Mahboob, K., Asif, R., & Haider, N. G. (2024). Career planning matters: Intelligence-based career path predictions using data mining models - A longitudinal study. *Mehran University Research Journal of Engineering and Technology*, *43*(4). https://doi.org/10.22581/muet1982.3343

Morales, J. (2017). La orientación vocacional para la elección de carreras universitarias dirigida a estudiantes de educación media. *Revista Internacional de Investigación y Formación Educativa*, 39–76. https://www.ensj.edu.mx/wp-content/uploads/2017/09/La-orientaci%C3%B3n-vocacional-para-la-elecci%C3%B3n.pdf

Mumme, C., Leipert, L. M., & Vollmeyer, R. (2025). Motivational reasons for dropping out of a physics degree program and gender differences in expectancies and values. *Discover Education 2025 4:1*, *4*(1), 54-. https://doi.org/10.1007/S44217-025-00442-5

Navarro Guzmán, C., & Casero Martinez, A. (2012). Análisis de las diferencias de género en la elección de estudios universitarios. Analysis of Gender Differences in Degree Choice. *ESTUDIOS SOBRE EDUCACIÓN*, *22*, 115–132. https://scispace.com/pdf/analisis-de-las-diferencias-de-genero-en-la-eleccion-de-4yfc4yynzw.pdf

Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., & 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

O'neill, K. (2024). *Predicting First Year Retention for Undergraduate Educational Opportunity Fund Students*. https://www.ramapo.edu/dmc/wp-content/uploads/sites/361/2024/05/msam1.pdf

Oqaidi, K., Aouhassi, S., & Mansouri, K. (2022). Towards a Students' Dropout Prediction Model in Higher Education Institutions Using Machine Learning Algorithms. *International Journal of Emerging Technologies in Learning*, *17*(18). https://doi.org/10.3991/ijet.v17i18.25567

Orozco-Rodríguez, C., Viegas, C., Costa, A. R., Lima, N., & Alves, G. R. (2025). Dropout Rate Model Analysis at an Engineering School. *Education Sciences*, *15*(3). https://doi.org/10.3390/EDUCSCI15030287

Rebelo Marcolino, M., Reis Porto, T., Thompsen Primo, T., Targino, R., Ramos, V., Marques Queiroga, E., Munoz, R., & Cechinel, C. (2025). Student dropout prediction through machine learning optimization: insights from moodle log data. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-93918-1

Rodriguez Ramirez, J. A., García-Bedoya, O., & Galpin, I. (2022). Maximizing Student Retention using Supervised Models Informed by Student Counseling Data. *In ICAI Workshops*, 225–239. https://ceur-ws.org/Vol-3282/icaiw_wdea_1.pdf

Rodríguez-Muñiz, L. J., Areces, D., Suárez-Álvarez, J., Cueli, M., & Muñiz, J. (2019). ¿Qué motivos tienen los estudiantes de Bachillerato para elegir una carrera universitaria? What motives have high school students for choosing a college degree? *Revista de Psicología y Educación / Journal of Psychology and Education*, 1–15. https://doi.org/10.23923/rpye2019.01.167

Santoso, H. B. (2020). Fuzzy Decision Tree to Predict Student Success in Their Studies. *International Journal of Quantitative Research and Modeling*, *1*(3). https://doi.org/10.46336/ijqrm.v1i3.59

Savickas, M. L. (2019). Career counseling (2nd ed.). In *Career counseling (2nd ed.).* American Psychological Association. https://doi.org/10.1037/0000105-000

Schlegel, B. E. (2026). *Logistic Regression*. 257–269. https://doi.org/10.1007/978-3-658-49801-6_19

Suárez-Perdomo, A., Álvarez-Pérez, P. R., & López-Aguilar, D. (2025). Una Revisión Sistemática sobre el Problema del Abandono Académico Universitario. *Electronic Journal of Research in Educational Psychology*, *23*(66). https://doi.org/10.25115/ejrep.v23i66.10500

Tinto, V. (1993). *Leaving College: Rethinking the Causes and Cures of Student Attrition, Tinto* (2d ed.). The University of Chicago Press. https://api.pageplace.de/preview/DT0400.9780226922461_A32289663/preview-9780226922461_A32289663.pdf

Tsai, S. C., Chen, C. H., Shiao, Y. T., Ciou, J. S., & Wu, T. N. (2020). Precision education with statistical learning and deep learning: a case study in Taiwan. *International Journal of Educational Technology in Higher Education*, *17*(1). https://doi.org/10.1186/s41239-020-00186-2

Whiston, S. C., Li, Y., Goodrich Mitts, N., & Wright, L. (2017). Effectiveness of career choice interventions: A meta-analytic replication and extension. *Journal of Vocational Behavior*, *100*, 175–184. https://doi.org/10.1016/j.jvb.2017.03.010

Zhu, F., Cai, Z., Buchtel, E. E., & Guan, Y. (2019). Career construction in social exchange: a dual-path model linking career adaptability to turnover intention. *Journal of Vocational Behavior*, *112*, 282–293. https://doi.org/10.1016/J.JVB.2019.04.003

Apéndices

Apéndice A

Diagrama de Flujo PRISMA 2020

Título de la revisión: Modelos predictivos en la orientación vocacional

Fase	Descripción	Cantidad
Identificación	Registros identificados mediante búsqueda en bases de datos (Google Scholar, Scopus, IEEE Xplore, Repositorio UNAD)	187
Eliminación de duplicados	Registros duplicados eliminados	43
	Registros tras eliminar duplicados	144
Cribado (título y resumen)	Registros excluidos por no cumplir criterios temáticos (no abordaban deserción, orientación vocacional o modelos predictivos supervisados)	82
	Registros evaluados en texto	62
Elegibilidad	Registros excluidos tras evaluar texto por: - Falta de claridad metodológica (12) - No incluir variables vocacionales ni de deserción (9) - Ser meramente descriptivos sin implementación de modelo (6)	27
Inclusión	Estudios incluidos en la revisión sistemática	35

Nota. Adaptado del modelo PRISMA para revisiones sistemáticas cualitativas. El criterio de inclusión ampliado abarcó publicaciones entre 2009 y 2026, en inglés o español, que presentaran modelos predictivos supervisados o revisiones sistemáticas/metaanálisis sobre deserción

universitaria u orientación vocacional. También se incluyeron estudios contextuales sobre la UNAD y reflexiones éticas sobre Big Data.

Apéndice B

Estudios Incluidos en la Revisión Sistemática

N°	Autor(es) y año	Título completo	Técnica(s) principal(es)	Variables principales	Métricas reportadas
1	Heredia et al., 2015	Student Dropout Predictive Model Using Data Mining Techniques	Árboles de decisión (C4.5, ID3), Regresión logística	Datos personales, socioeconómicos, desempeño académico	Precisión, comparación con SPADIES
2	Ramírez & Grandón, 2018	Prediction of student dropout in a Chilean public university through classification based on decision trees with optimized parameters	Árboles de decisión (CART) con optimización de parámetros	Notas, asistencia, variables demográficas	Precisión (87,27%)
3	Avila Pérez & Medina, 2020	Minería de datos para la predicción de la deserción estudiantil en la Universidad Nacional Abierta y a Distancia	Árboles de decisión	Datos de caracterización de estudiantes UNAD, acceso a plataforma	Precisión, sensibilidad
4	Avila Pérez,	Modelo de predicción de	Árboles de	Accesos al campus	Precisión

	2021	deserción estudiantil, apoyado en Tecnologías De Data Mining, en un curso de primera matrícula de la Escuela ECBTI De La UNAD	decisión, Random Forest	virtual, entrega de actividades, participación en foros	(~74%)
5	Kemper et al., 2020	Predicting student dropout: A machine learning approach	AdaBoost, redes neuronales, árboles de decisión	Datos administrativos, rendimiento académico previo, asistencia	Exactitud, AUC-ROC
6	Tsai et al., 2020	Precision education with statistical learning and deep learning: a case study in Taiwan	Deep learning, regresión logística	Ranking de calificaciones, solicitudes de préstamo estudiantil, ausencias, número de asignaturas con alerta	Exactitud (77% deep learning, 68% estadístico), sensibilidad, especificidad
7	Dake & Buabeng-Andoh, 2022	Using Machine Learning Techniques to Predict Learner Drop-out Rate in Higher Educational Institutions	Random Forest, SVM, MLP, árboles de decisión	Rendimiento académico previo, asistencia, nivel socioeconómico	Exactitud (70,98% Random Forest), precisión, recall, F1-score, ROC
8	Rodriguez Ramirez et al., 2022	Maximizing Student Retention using Supervised Models Informed by Student	Random Forest, XGBoost	Variables académicas, socioeconómicas, datos de asesoría estudiantil	AUC-ROC (>0,85), F1-score (0,81),

		Counseling Data		(indecisión, satisfacción)	precisión
9	Niyogisubizo et al., 2022	Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization	Ensemble (Random Forest, XGBoost, Gradient Boosting, FNN)	Datos académicos, demográficos, historial de cursos	Exactitud, AUC-ROC
10	Oqaidi et al., 2022	Towards a Students' Dropout Prediction Model in Higher Education Institutions Using Machine Learning Algorithms	Revisión de múltiples algoritmos	Síntesis de variables académicas, socioeconómicas, institucionales	Comparación de métricas (exactitud, AUC, F1)
11	Calero et al., 2023	Factores determinantes de la deserción escolar en la región Huánuco, Perú	Regresión logística	Nivel socioeconómico, familia monoparental, padres desertores, edad, sexo, necesidad de trabajar	Odds ratios, probabilidades
12	Eegdeman, 2023	Enhancing Study Success in Dutch Vocational Education (tesis)	Random Forest, XGBoost	Autoeficacia vocacional, motivación, rendimiento previo, asistencia	AUC-ROC (0,78-0,82), sensibilidad, F1-score
13	Benoit et al., 2024	High-stake student drop-out prediction using hidden Markov models in fully	Hidden Markov Models (HMM)	Comportamiento de aprendizaje en MOOC, motivación, interacciones	Exactitud, sensibilidad, especificidad

		asynchronous subscription-based MOOCs			
14	Harnisher et al., 2024	Collaborative Development of Machine Learning Algorithms for Student Success at John Jay College	Machine learning colaborativo (varios algoritmos)	Datos académicos, asesoría, intervenciones	Graduación, retención (métricas cualitativas)
15	Loder, 2024	Multiple Enrollment Policy: Survival Analyses and Odds of Graduating in at Least One University Degree Program	Regresión logística, análisis de supervivencia	Múltiples matriculaciones simultáneas, pre-estudios, transferencia de créditos	Odds ratios, curvas de supervivencia
16	O'Neill, 2024	Predicting First Year Retention for Undergraduate Educational Opportunity Fund Students	Random Forest con SMOTE	GPA, variables socioeconómicas, pertenencia a programa EOF	Exactitud, F1-score, AUC-ROC
17	Mahboob et al., 2024	Career planning matters: Intelligence-based career path predictions using data mining models - A longitudinal study	Múltiples modelos (árboles, regresión logística)	Rendimiento académico (CGPA, FYP), estatus económico, demografía, educación preuniversitaria	Precisión, recall, F1-score
18	Zhu et al., 2019	Career construction in social exchange: a dual-path model linking career adaptability to	Regresión logística (turnover	Adaptabilidad profesional, satisfacción con la carrera, apoyo	Razones de probabilidad, efectos de

		turnover intention	intention)	organizacional percibido	mediación
19	Orozco-Rodríguez et al., 2025	Dropout Rate Model Analysis at an Engineering School	Regresión logística	Género, desplazamiento del hogar, promedio escolar en secundaria, habilidades matemáticas	Odds ratios, significación estadística
20	Huamán Arratia, 2025	Modelo predictivo para la prevención de la deserción académica en estudiantes de un instituto tecnológico de Puno, 2025	Regresión logística, árboles de decisión	Rendimiento primer semestre, condiciones socioeconómicas, edad, género	Precisión, recall (sensibilidad)
21	Muñoz de Luna & Martin Gomez, 2025	APLICACIÓN DE LA REGRESIÓN LOGÍSTICA BINARIA EN LA EDUCACIÓN ASISTIDA POR INTELIGENCIA ARTIFICIAL	Regresión logística	Competencias digitales, uso de ChatGPT, edad, área de estudio	Odds ratios, probabilidades
22	Rebelo Marcolino et al., 2025	Student dropout prediction through machine learning optimization: insights from moodle log data	CatBoost con balanceo Adaptive Synthetic	Registros de actividad en Moodle (logs), patrones de interacción	F1-score (~0,8), recall
23	Gochhayat & Ravindran, 2025	Drop out v/s retention of female students: unfolding dynamics of the education system in Indian states	Regresión logística multinomial	Género, edad, nivel educativo de los padres, rural/urbano	Odds ratios, significación

24	Suárez-Perdomo et al., 2025	Una Revisión Sistemática sobre el Problema del Abandono Académico Universitario	Revisión sistemática	Síntesis de variables individuales (autoconfianza, compromiso, motivación), académicas, socioeconómicas, institucionales	No aplica (síntesis cualitativa)
25	Kocsis & Molnár, 2025	Factors influencing academic performance and dropout rates in higher education	Revisión (EDM, regresión logística, redes neuronales)	GPA, créditos ECTS, género, motivación, autoeficacia, trabajo, finanzas	Exactitud, AUC (comparativa)
26	García-Botero et al., 2022	REVISIÓN SISTEMÁTICA: CRITERIOS DE CALIDAD EN EL PROYECTO DE PROGRAMAS DE ORIENTACIÓN VOCACIONAL	Revisión sistemática	Intereses, aptitudes, autoeficacia, indecisión, uso de TICs, rol del orientador	No aplica
27	Whiston et al., 2017	Effectiveness of career choice interventions: A meta-analytic replication and extension	Metaanálisis	Autoeficacia para la toma de decisiones, indecisión vocacional, identidad vocacional	Tamaño del efecto (d de Cohen)
28	Cabus & De Witte, 2016	Why Do Students Leave Education Early? Theory and Evidence on High School	Modelo económico predictivo	Preferencias temporales, motivación, costos de oportunidad, políticas	Probabilidades, tasas de retención

		Dropout Rates		preventivas	
29	Mumme et al., 2025	Motivational reasons for dropping out of a physics degree program and gender differences in expectancies and values	Regresión logística	Autoconcepto académico, valor utilitario, valor intrínseco, costos psicológicos, género	Razones de probabilidad
30	Coetzee et al., 2023	To what extent do career agility and psychological capital activate employees' career adaptability and foster their career resilience and career satisfaction?	Regresión (mediación)	Agilidad profesional, capital psicológico, adaptabilidad profesional, resiliencia, satisfacción	Coeficientes de mediación, efectos directos
31	Loder, 2025	Machine learning for university management: Micro Cluster Learning to predict "active" students	Micro Cluster Learning (híbrido de ML y ARIMA)	Datos administrativos universitarios, trayectorias históricas	Desviación vs. estadísticas oficiales (2-8%)
32	Federici et al., 2021	The moderating role of HR practices on the career adaptability–job crafting relationship: a study among employee–manager dyads	Regresión (moderación)	Adaptabilidad profesional, job crafting, prácticas de alto rendimiento, engagement	Efectos de moderación, significación
33	Cedillo-Quizhpe et	Influencia de las expectativas académicas sobre los	Ecuaciones estructurales	Expectativas académicas (desarrollo personal,	Coeficientes estructurales, R^2

	al., 2026	motivos de intención de abandono en estudiantes de una universidad ecuatoriana	(PLS-SEM)	interacción social, implicación política), opción de carrera, nivel educativo de los padres	
34	Johnson, 2014	The Ethics of Big Data in Higher Education	Reflexión ética (no predictivo)	Privacidad, consentimiento informado, sesgo algorítmico, integridad contextual	No aplica (marco ético)
35	Facundo Díaz, 2009	Análisis sobre la deserción en la educación superior a distancia y virtual: el caso de la UNAD - COLOMBIA	Análisis contextual (no predictivo)	Factores institucionales, conectividad, compatibilidad estudio-trabajo, integración virtual	Tasas de deserción descriptivas

Nota. Este listado incluye los 35 estudios que conforman la muestra final de la revisión sistemática, de acuerdo con los criterios ampliados (2009-2026, modelos predictivos supervisados o contextuales relevantes). Los estudios 34 y 35 no presentan modelos predictivos formales, pero se incluyen por su valor contextual y ético para la discusión.