MODELO DE PREDICCIÓN DE DESERCIÓN EN ESTUDIANTES DE UNA INSTITUCIÓN DE EDUCACIÓN SUPERIOR

José Julián Figueroa Arias
Código 8993441
Yudy Alexandra Porras Ríos
Código 8979818

Director
Cristian Alejandro Torres Valencia

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, 25 DE MAYO DE 2026

FICHA RESUMEN
TRABAJO DE GRADO DE MAESTRÍA

Título del proyecto: Modelo de predicción de deserción en estudiantes de una institución de educación superior

1. ÁREA DE TRABAJO: Educación
2. TIPO DE PROYECTO: Aplicado
3. ESTUDIANTES: Yudy Alexandra Porras Ríos y José Julián Figueroa Arias
4. CORREO ELECTRÓNICO: julianfigueroa26@javerianacali.edu.co
5. DIRECCIÓN Y TELEFONO: Calle 8 no 20 a 23 (Pasto, N) – 3188682802-3127042546
6. DIRECTOR: Cristian Alejandro Torres Valencia
7. VINCULACIÓN DEL DIRECTOR: Docente PUJ Cali
8. CORREO ELECTRÓNICO DEL DIRECTOR: cristian.torres@javerianacali.edu.co
9. CODIRECTOR: No Aplica
10. GRUPO O EMPRESA QUE LO AVALA: Datos Abiertos
11. OTROS GRUPOS O EMPRESAS: No Aplica
12. PALABRAS CLAVE: Machine Learning, Deserción Universitaria, Modelo Predictivo.
13. FECHA DE INICIO: 15/05/2025
14. FECHA DE FINALIZACIÓN: 25/05/2026
15. RESUMEN: La deserción estudiantil constituye una de las principales problemáticas que enfrentan las instituciones de educación superior. En Colombia, aproximadamente el 36% de los estudiantes abandona sus estudios durante el primer año, lo que resalta la importancia de abordar este fenómeno mediante herramientas analíticas. En este contexto, el presente trabajo tuvo como objetivo desarrollar un modelo de predicción de deserción estudiantil utilizando técnicas de aprendizaje automático supervisado, a partir del conjunto de datos público "Predict Students' Dropout and Academic Success" del UCI Machine Learning Repository, correspondiente a una institución de educación superior portuguesa con 4.424 registros. La metodología se estructuró bajo el estándar CRISP-DM, abarcando las fases de comprensión y preparación de los datos, modelado y evaluación. Se entrenaron y compararon cinco algoritmos: regresión logística, árbol de decisión, Random Forest, XGBoost y redes neuronales artificiales, bajo un esquema de validación cruzada estratificada repetida. Los resultados evidenciaron que XGBoost presentó el mejor desempeño global en la identificación del riesgo, alcanzando un Recall de 0.9085, un ROC-AUC de 0.9728 y un F1-score de 0.8990 en el conjunto de prueba. Adicionalmente, el modelo seleccionado fue integrado en un prototipo funcional desarrollado con Streamlit y desplegado en la nube, que permite a los equipos institucionales visualizar las predicciones de riesgo de forma interactiva y tomar decisiones de intervención basadas en datos.

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE ANEXOS

INTRODUCCIÓN

La deserción estudiantil en instituciones de educación superior constituye una problemática estructural que impacta significativamente la calidad académica, la eficiencia institucional y el desarrollo tanto económico como social de los países. En Colombia, se estima que cerca del 36 % de los estudiantes abandonan sus estudios durante el primer año, cifra que revela la necesidad de fortalecer los mecanismos de prevención y acompañamiento que favorezcan la permanencia estudiantil [1]. Esta situación, además de generar pérdidas económicas para las instituciones, limita la formación de capital humano y profundiza desigualdades sociales, especialmente en poblaciones vulnerables.

En ese sentido, la incorporación de técnicas de ciencia de datos y aprendizaje automático ha emergido como una alternativa metodológica robusta para el análisis y predicción de la deserción estudiantil. Diversos estudios han demostrado que los modelos predictivos permiten identificar patrones complejos a partir de datos académicos, sociodemográficos y económicos, facilitando la detección temprana de estudiantes en riesgo y la implementación de intervenciones oportunas [2]. Sin embargo, muchas instituciones de educación superior aún carecen de herramientas analíticas integradas que permitan aprovechar de manera sistemática la información disponible para la toma de decisiones.

En respuesta a esta necesidad, el proyecto actual tiene como objetivo desarrollar un modelo de predicción de deserción estudiantil mediante técnicas de aprendizaje automático supervisado. Para ello, se emplea el conjunto de datos “Predict Students’ Dropout and Academic Success” del repositorio UCI; el cual contiene información histórica de 4424 estudiantes. La metodología adoptada se basó en el estándar CRISP DM, abarcando desde la comprensión del problema hasta la evaluación comparativa de los modelos.

Se implementaron y evaluaron diferentes algoritmos de clasificación, incluyendo regresión logística, árboles de decisión, bosques aleatorios (*random forest*), XGBoost y redes neuronales artificiales; utilizando métricas como accuracy, precision, recall, F1-score y ROC-AUC. Los resultados evidencian que los modelos de ensamble y aprendizaje profundo presentan un desempeño superior en la identificación de estudiantes en riesgo, destacándose XGBoost por su alta capacidad discriminativa.

Por último, el proyecto incluye el desarrollo de un prototipo funcional que permite visualizar las predicciones del modelo seleccionado mediante un dashboard interactivo, facilitando su interpretación por parte de actores institucionales. De esta manera, se busca contribuir al diseño de estrategias basadas en datos para la reducción de la deserción estudiantil en instituciones de educación superior.

1. DEFINICIÓN DEL PROBLEMA

La deserción estudiantil representa una de las principales preocupaciones en las instituciones de educación superior (IES), al afectar no sólo los indicadores de calidad y eficiencia del sistema educativo; sino también el desarrollo social y económico de los países [3],[4]. Un estudio del año 2021 [5], que evalúa los indicadores de deserción universitaria; enuncia que las tasas de abandono varían según el contexto geográfico, denotando en la Unión Europea cifras que oscilan entre el 20% al 55%, mientras que en América Latina las tasas fluctúan entre un 8 a 48%. Por su parte, una investigación de la Universidad del Bosque reporta que en Colombia el 36% de los estudiantes se retiran durante el primer año [1]. Llama la atención, estos resultados y más atendiendo al grado de evolución que presentan los sistemas educativos en países desarrollados. Sin embargo, estas cifras pueden estar influenciadas por la limitada disponibilidad de datos en las instituciones educativas de Latinoamérica.

1.1. PLANTEAMIENTO DEL PROBLEMA

Frente a esta problemática, el uso de modelos de predicción basados en ciencia de datos surge como una herramienta útil e innovadora para anticipar comportamientos de abandono y fortalecer las estrategias de permanencia académica. Pues, mediante el análisis de grandes volúmenes de datos, en donde se incluyan variables sociodemográficas, académicas, psicosociales y contextuales, es posible detectar patrones ocultos que explican la deserción [6],[7],[8]. No obstante, es menester resaltar que a la fecha diferentes instituciones educativas aún carecen de modelos propios que integren de manera efectiva estas herramientas, lo que limita su capacidad para implementar estrategias tempranas de retención.

En consecuencia, las IES enfrentan una débil capacidad para mitigar y anticipar la deserción estudiantil; debido a la ausencia de estrategias claras y articuladas de seguimiento académico y personal. Si bien, existen programas de bienestar que abordan señales tempranas como inasistencia, el bajo rendimiento o el retraso en pagos; la información recolectada suele estar fragmentada, sin integrarse en sistemas que permitan un análisis riguroso. Esta dispersión dificulta la identificación oportuna de patrones de abandono y limitan la implementación de intervenciones preventivas [9].

Por tanto, las acciones institucionales son reactivas, mínimamente personalizadas y con bajo impacto en la permanencia estudiantil. En donde, la falta de mecanismos sistemáticos de monitoreo, análisis y gestión de datos representa una barrera crítica para abordar de manera efectiva la deserción en la educación superior.

1.2. FORMULACIÓN DEL PROBLEMA

Al reconocer, que la deserción académica conlleva impactos negativos duraderos, tales como menores ganancias a lo largo de la vida, dificultades para ingresar al mercado laboral formal, entre otros [10], se da lugar a la siguiente pregunta de investigación:

¿Cómo predecir la deserción estudiantil en instituciones de educación superior a través de técnicas de aprendizaje automático supervisado?

2. OBJETIVOS

2.1. OBJETIVO GENERAL

Desarrollar un modelo de predicción de deserción en estudiantes de una institución de educación superior a través de técnicas de aprendizaje automático supervisado

2.2. OBJETIVOS ESPECÍFICOS

- Preparar el conjunto de datos para el entrenamiento y prueba de los modelos, mediante procesos de limpieza, transformación y selección de variables de tipo socioeconómicas, académicas y personales.
- Implementar algoritmos de aprendizaje automático supervisado para realizar la predicción de deserción estudiantil en una institución de educación superior.
- Evaluar el rendimiento de los modelos a través de métricas de desempeño para algoritmos de aprendizaje automático supervisado, identificando el modelo con mejor desempeño en la predicción de deserción estudiantil.
- Implementar un prototipo funcional que permita visualizar las predicciones realizadas por el mejor modelo.

3. MARCO TEÓRICO Y ANTECEDENTES

El estudio de la deserción estudiantil en educación superior requiere un abordaje integral que articule perspectivas teóricas del ámbito educativo con enfoques analíticos provenientes de la ciencia de datos. En este sentido, el presente marco integra dos dimensiones fundamentales, por una parte, la comprensión del fenómeno de la deserción y por otra, la revisión de metodologías de aprendizaje automático supervisado utilizadas en la construcción de modelos predictivos.

3.1. MARCO TEÓRICO

3.1.1. Deserción en universidades

La deserción universitaria hace referencia al abandono de los estudios antes de finalizar un programa académico. Este fenómeno es resultado de factores, como problemas

económicos, falta de motivación, bajo desempeño académico, dificultades de adaptación al entorno universitario y una oferta insuficiente de apoyo por parte de las instituciones [11]. Además, este problema no solo impacta a los estudiantes que interrumpen su formación, sino que también afecta a las instituciones, reduciendo las tasas de graduación y desaprovechando los recursos invertidos. En términos sociales, la deserción implica una pérdida de capital humano que podría contribuir al desarrollo económico y social.

Para los estudiantes, abandonar la universidad significa la pérdida de oportunidades de desarrollo personal y profesional, lo que generalmente se traduce en una menor capacidad para acceder a empleo [12]. La deserción estudiantil perpetúa las desigualdades sociales, ya que muchos jóvenes de comunidades vulnerables se ven obligados a abandonar sus estudios debido a la falta de recursos, lo que limita su movilidad y las posibilidades de mejorar su calidad de vida.

Por otro lado, la deserción impacta negativamente en la estabilidad financiera de las universidades. Las altas tasas de deserción afectan la proyección de los planes académicos y administrativos, reduciendo los ingresos por matrícula. Por su parte a nivel social la deserción universitaria tiene un impacto directo en la formación del capital humano, crucial para enfrentar los desafíos globales y mejorar la productividad de un País. Esta falta de profesionales contribuye a la escasez de talento en sectores estratégicos, lo que frena la capacidad para generar nuevos proyectos empresariales y mejorar sus estándares laborales [13].

Ahora bien, teniendo como referencia a Castaño [11] existen las siguientes categorías de factores que determinan la deserción en estudiantes universitarios.

- Factores individuales: en este caso se encuentran variables tales como la edad, sexo, género, estado civil, situación de salud
- Factores académicos: satisfacción, calidad académica, modalidad de estudio, acompañamiento estudiantil.
- Factores institucionales: acceso a becas, acceso a financiamiento, modalidades de estudio, interacción entre docente y estudiante, recursos de la universidad.
- Factores socioeconómicos: situación laboral, nivel de ingresos del, situación laboral de los padres, nivel de ingresos de los padres, nivel educativo de los padres, financiamiento de estudios, personas a cargo.

3.1.2. Modelos de aprendizaje automático supervisado

La predicción de la deserción estudiantil constituye un problema de clasificación binaria en el cual se busca estimar la probabilidad de que un estudiante pertenezca a una de dos clases: permanencia o abandono [14]. En este contexto, los algoritmos de aprendizaje automático

supervisado permiten construir modelos predictivos capaces de identificar patrones complejos en datos históricos académicos, socioeconómicos y administrativos para anticipar comportamientos futuros [14]. Esta aproximación basada en datos facilita la detección temprana de riesgos, posibilitando intervenciones proactivas a nivel institucional.

Diversos estudios han demostrado que los modelos predictivos fundamentados en técnicas de ciencia de datos contribuyen significativamente a mejorar la identificación de estudiantes en riesgo de abandono [2], [15]. Esta capacidad predictiva permite a las instituciones educativas diseñar e implementar estrategias de intervención oportunas y personalizadas, orientadas específicamente a fortalecer la permanencia académica de los estudiantes vulnerables [14]. La precisión en la predicción temprana es crítica para optimizar la asignación de recursos institucionales destinados a retención estudiantil.

En el presente estudio se evaluaron diversos algoritmos de aprendizaje supervisado con diferentes niveles de complejidad estructural, con el propósito de comparar su desempeño predictivo en la identificación de estudiantes con riesgo de deserción [15]. Los modelos evaluados abarcan enfoques estadísticos tradicionales, métodos basados en árboles de decisión, algoritmos de ensamble y arquitecturas de redes neuronales. Esta estrategia comparativa permite analizar el compromiso entre complejidad del modelo e interpretabilidad, aspectos fundamentales en contextos educativos donde explicar las predicciones es esencial para la toma de decisiones institucionales [15].

3.1.2.1. Regresión Logística

La regresión logística es un modelo ampliamente utilizado en problemas de clasificación binaria [16]. Su objetivo fundamental consiste en estimar la probabilidad de ocurrencia de un evento de interés a partir de un conjunto de variables explicativas [16]. A diferencia de la regresión lineal tradicional, la regresión logística emplea una función logística o sigmoide que restringe las predicciones al intervalo (0, 1), lo que permite interpretaciones probabilísticas válidas [14]. La forma funcional del modelo de regresión logística se expresa mediante la siguiente ecuación:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)}}$$

donde P (Y = 1|X) representa la probabilidad de que la variable dependiente Y tome el valor 1 (evento de interés) dado un conjunto de variables independientes X, e es la base del logaritmo natural, y β_o, β_1, ..., β_p son los coeficientes del modelo a estimar [16]. Esta función de transformación logística garantiza que las predicciones se encuentren siempre en el rango de probabilidad válido (0, 1).

Alternativamente, el modelo puede expresarse mediante la función logit, que representa el logaritmo de la razón de probabilidades (*odds ratio*):

$$logit(P) = \ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Esta transformación linealiza la relación entre las variables independientes y el *log-odds* de la variable dependiente, facilitando la interpretación de los coeficientes [14].

La regresión logística ha sido ampliamente utilizada en diversos contextos de predicción binaria debido a su robustez, eficiencia computacional e interpretabilidad [17]. En estudios comparativos de modelos predictivos, se utiliza frecuentemente como modelo de referencia o baseline, permitiendo evaluar el desempeño relativo de algoritmos más complejos respecto a esta línea base [17]. La capacidad de la regresión logística para proporcionar no solo predicciones de clases sino también probabilidades asociadas, la convierten en una herramienta valiosa para la toma de decisiones en contextos donde la cuantificación de la incertidumbre es crítica [16]. Su simplicidad interpretativa facilita la comunicación de resultados a audiencias no especializadas en aprendizaje automático.

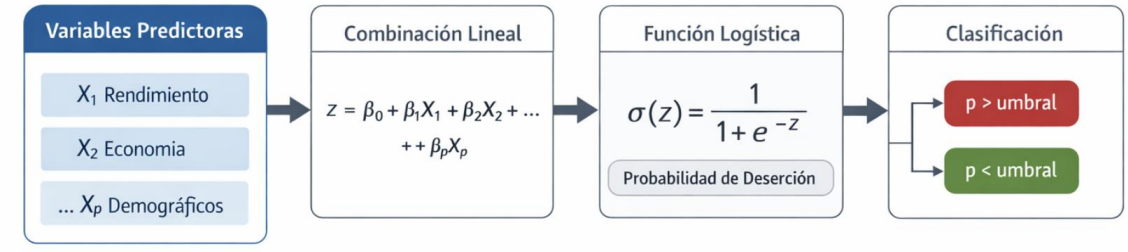


Figura 1. Esquema regresión logística

3.1.2.2. Árboles de Decisión

Los árboles de decisión son modelos jerárquicos utilizados para representar decisiones y sus posibles resultados mediante una estructura intuitiva de nodos y ramas [18]. Cada nodo interno representa una prueba basada en un atributo específico, mientras que los nodos hoja indican los resultados finales predichos. Esta arquitectura permite descomponer problemas complejos en una serie de decisiones binarias secuenciales. Cada nodo del árbol plantea una regla de decisión de la forma:

$$\text{Si } X_j \leq \theta \text{ entonces rama izquierda, sino rama derecha}$$

donde X_j representa la variable predictora seleccionada en el nodo actual y θ denota el punto de corte o umbral de división. A partir de esta regla, los datos se dividen en dos subconjuntos: la rama izquierda contiene las instancias que satisfacen la condición ($X_j \leq \theta$),

mientras que la rama derecha agrupa aquellas que no la satisfacen. Este proceso se repite recursivamente en cada nodo hasta alcanzar criterios de parada predefinidos.

Este enfoque es ampliamente usado en inteligencia artificial, estadística y análisis de datos gracias a su capacidad para manejar relaciones complejas no lineales [19]. La principal ventaja radica en su interpretabilidad visual, que permite identificar patrones y predecir resultados de forma intuitiva, siendo especialmente útil en áreas como segmentación de clientes y proyección de ventas [20]. La transparencia inherente de estos modelos facilita la comunicación de resultados a audiencias no especializadas.

Sin embargo, estos modelos pueden presentar sobreajuste cuando enfrentan datos ruidosos o realizan demasiadas divisiones, comprometiendo su capacidad de generalización [18]. Para mitigar este problema, es necesario aplicar técnicas de poda que reducen la complejidad del árbol. La construcción óptima de árboles se basa en algoritmos como ID3, C4.5 y CART, que seleccionan divisiones mediante métricas de impureza tales como la ganancia de información o el índice de Gini [20]. Estas métricas cuantifican la reducción de incertidumbre lograda en cada partición de datos.

Adicionalmente, los árboles de decisión sirven como base conceptual para métodos más complejos y robustos, tales como bosques aleatorios y algoritmos de *boosting*, que combinan múltiples árboles para aumentar sustancialmente la precisión y estabilidad de las predicciones [19]. Esta versatilidad ha consolidado a los árboles de decisión como una herramienta fundamental tanto en tareas de clasificación como en problemas de regresión supervisada, representando un equilibrio óptimo entre simplicidad interpretativa y capacidad predictiva.

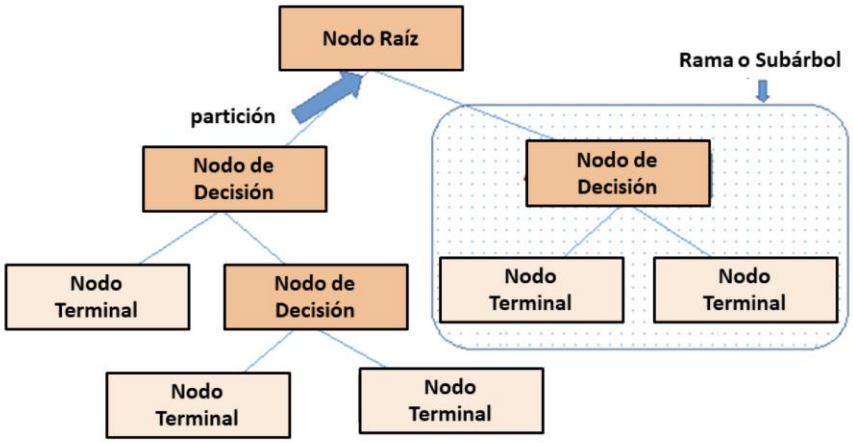


Figura 2. Diagrama árboles de decisión [21]

3.1.2.3. Bosques Aleatorios (Random Forest)

El algoritmo Random Forest es un método de ensamble basado en la técnica de *bagging* (*Bootstrap Aggregating*), fue propuesto por Breiman (2001) y su objetivo es mejorar la estabilidad y precisión de los árboles de decisión individuales [22]. Este modelo construye múltiples árboles de decisión utilizando diferentes muestras *bootstrap* del conjunto de datos y combina sus predicciones mediante votación mayoritaria en problemas de clasificación, o promedio en problemas de regresión [22]. El proceso de *bootstrap* permite que cada árbol sea entrenado con muestras diferentes, lo que reduce el sesgo individual y contribuye a la robustez del modelo final. Además, en cada división del árbol se selecciona aleatoriamente un subconjunto de variables predictoras, lo que introduce un grado adicional de aleatorización que reduce la correlación entre los árboles y mejora la capacidad de generalización del modelo [14].

El mecanismo de agregación de predicciones en Random Forest es fundamental para su efectividad. En problemas de clasificación, la predicción final se obtiene mediante votación mayoritaria: $\hat{y}\_RF = \text{mode}(\hat{y}_1, \hat{y}_2, ..., \hat{y}_n)$, donde cada $\hat{y}_i$ representa la predicción del árbol i-ésimo y mode representa la clase que recibe el mayor número de votos [22].

En problemas de regresión, la predicción final del modelo se calcula como el promedio de todas las predicciones individuales:

$$\widehat{y_{RF}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{y_i}$$

siendo n el número total de árboles en el bosque [23]. Esta estrategia de combinación simple pero poderosa reduce significativamente la varianza del modelo sin aumentar sustancialmente el sesgo, lo que resulta en un predictor más robusto. La naturaleza paralela del entrenamiento de árboles independientes también proporciona ventajas computacionales, permitiendo aprovechar arquitecturas de procesamiento paralelo para acelerar el proceso.

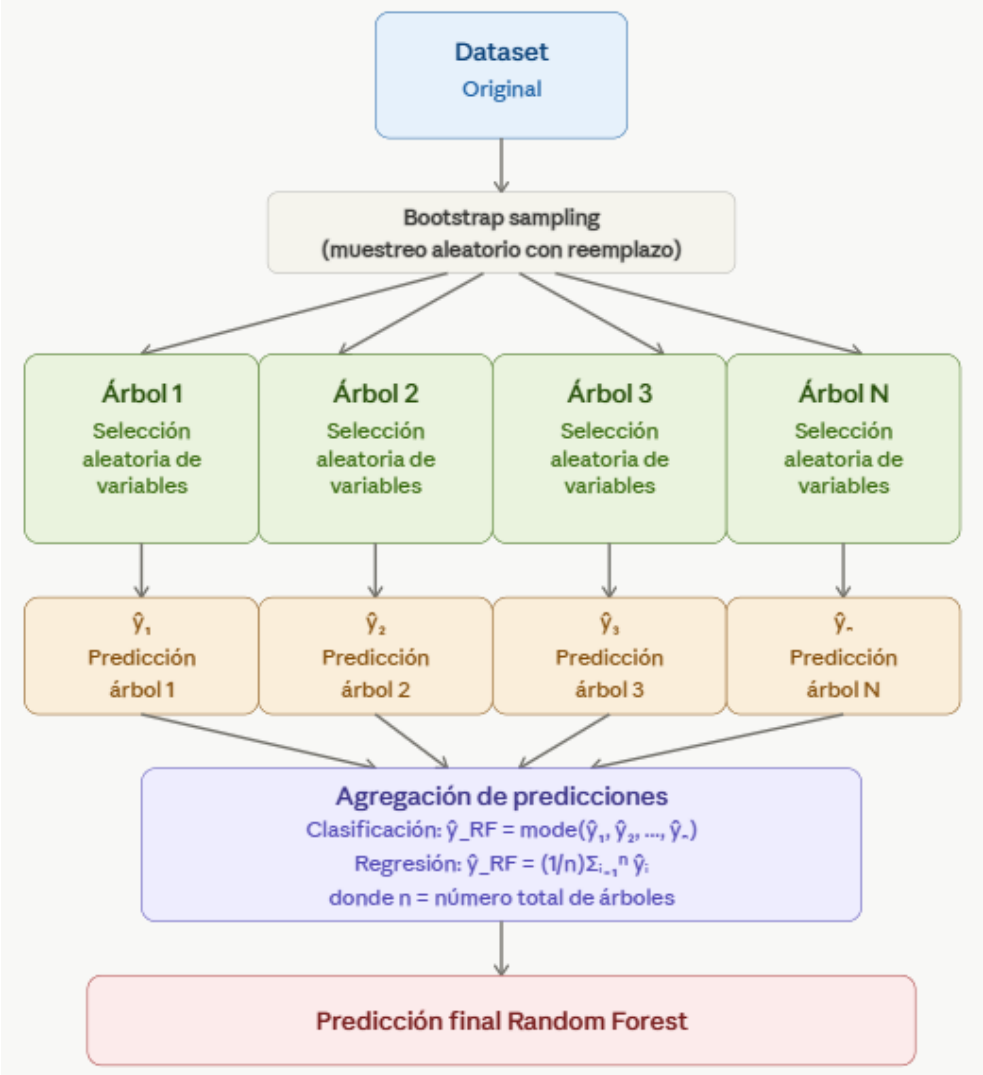


Figura 3. Proceso iterativo bosques aleatorios

3.1.2.4. Gradient Boosting y Xgboost

Los métodos de *gradient boosting* constituyen una familia de algoritmos de ensamble en los que múltiples modelos débiles, generalmente árboles de decisión poco profundos, se combinan secuencialmente mediante un proceso iterativo de aprendizaje [24]. A diferencia del *bagging*, donde los modelos se entrenan de forma independiente, en el *boosting* cada nuevo modelo se entrena específicamente para corregir los errores cometidos por los modelos anteriores. Este enfoque permite optimizar de manera progresiva una función de pérdida mediante técnicas de descenso por gradiente, lo que resulta en un predictor cada vez más robusto con cada iteración [24]. El principio fundamental es que la combinación ponderada de predictores débiles produce un modelo fuerte capaz de capturar relaciones complejas en los datos.

Extreme Gradient Boosting (XGBoost), propuesto por Chen y Guestrin (2016), representa una de las implementaciones más populares y eficientes de esta familia de algoritmos [25]. XGBoost realiza predicciones mediante una combinación aditiva de K árboles de decisión:

$$\hat{y} = \sum_{k=1}^{K} f_k(x)$$

donde cada función f_k representa un árbol de decisión individual. La optimización se realiza minimizando la siguiente función objetivo:

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

donde l es la función de pérdida, n es el número de muestras, e $\Omega(f_k)$ representa el término de regularización que penaliza la complejidad de cada árbol [25]. XGBoost incorpora varias mejoras significativas respecto a los métodos de *boosting* tradicionales: implementa regularización integrada que penaliza la complejidad de los árboles individuales para prevenir el sobreajuste, utiliza técnicas de optimización avanzadas que aceleran el entrenamiento, y permite manejar eficientemente valores faltantes sin requerimiento de imputación previa. Estas características lo hacen especialmente efectivo para trabajar con grandes conjuntos de datos estructurados, permitiendo que logre un balance óptimo entre sesgo y varianza.

En contextos educativos, XGBoost ha demostrado ser particularmente efectivo para tareas de predicción que involucran múltiples dimensiones de información [26]. Su capacidad para capturar interacciones no lineales entre variables, combinada con su robustez ante datos heterogéneos y su interpretabilidad relativa, permite identificar las variables de mayor importancia [26]. Además, su eficiencia computacional facilita la validación cruzada y ajuste de hiperparámetros, procesos críticos para garantizar que los modelos generalicen correctamente datos no vistos [27].

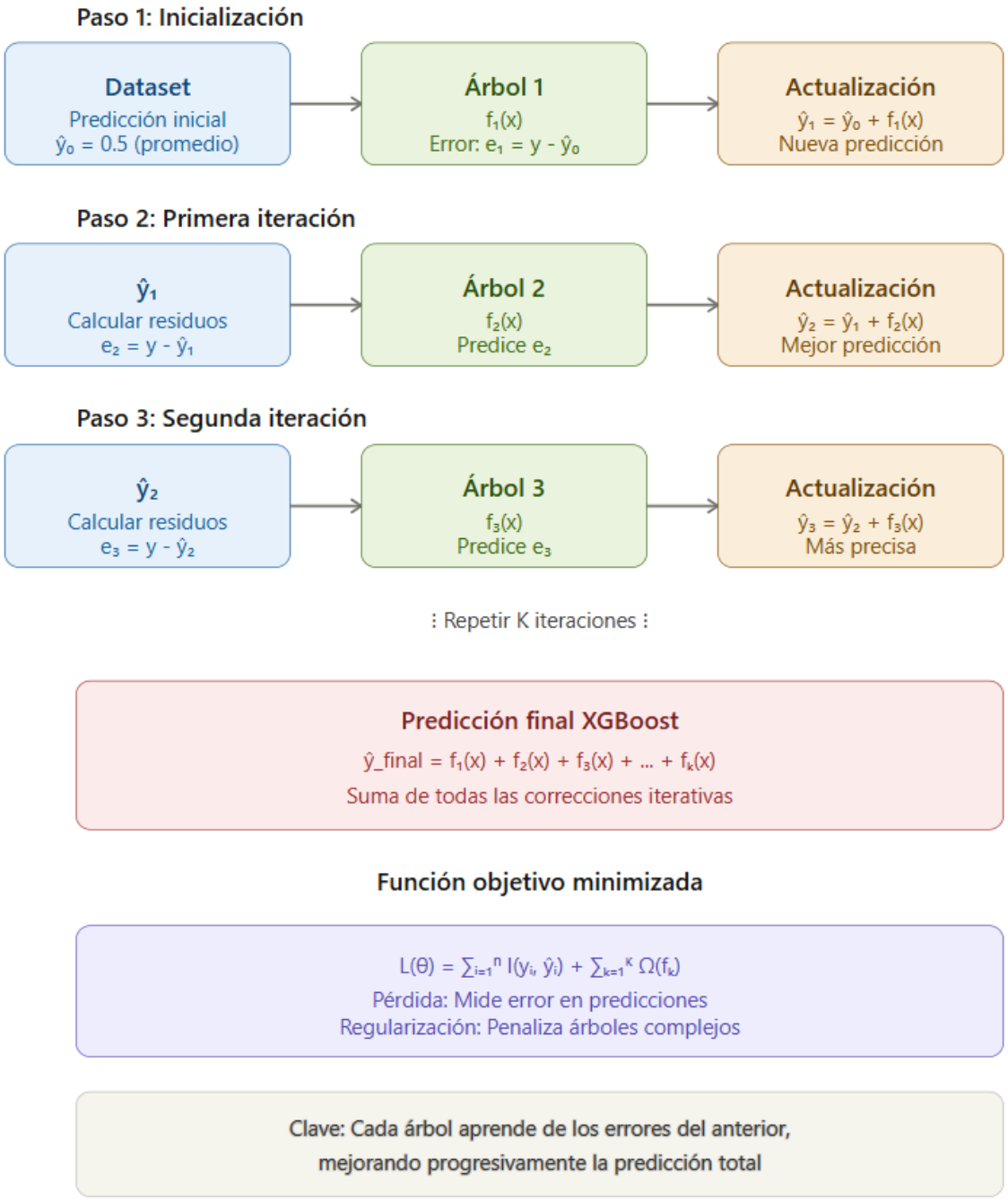


Figura 4. Proceso iterativo XGBOOST

3.1.2.5. Redes Neuronales (Fully Connected)

Las redes neuronales artificiales constituyen modelos computacionales inspirados en el funcionamiento del sistema nervioso biológico [28]. Estas redes están formadas por unidades de procesamiento denominadas neuronas artificiales, organizadas en capas que transforman progresivamente las variables de entrada mediante operaciones lineales y funciones de activación no lineales. Cada capa está conectada con todas las neuronas de la capa siguiente, conformando una arquitectura totalmente conectada (*fully connected*). En términos matemáticos, cada capa realiza una transformación de la forma:

$$h = f(Wx + b)$$

donde W representa la matriz de pesos, b el sesgo, x la entrada y f una función de activación no lineal como ReLU o Sigmoid [29]. Estas transformaciones sucesivas permiten a la red aprender representaciones jerárquicas complejas de los datos.

El proceso de entrenamiento de redes neuronales *fully connected* utiliza una metodología iterativa basada en el algoritmo de *backpropagation* [30]. Durante cada iteración (o época), la red realiza un *forward pass* donde los datos se propagan a través de todas las capas generando una predicción. Posteriormente, se calcula una función de pérdida que mide la discrepancia entre la predicción y el valor real. El componente crítico del entrenamiento es el *backward pass*, donde los gradientes de la función de pérdida se propagan hacia atrás a través de la red utilizando la regla de la cadena. Estos gradientes indican la dirección y magnitud en la que cada peso debe ajustarse para reducir el error. Finalmente, los pesos se actualizan mediante descenso de gradiente: w_nuevo = w_anterior - α × ∂L/∂w, donde α es la tasa de aprendizaje que controla la magnitud de cada ajuste. Este proceso se repite durante múltiples épocas hasta que la red converge a soluciones donde el error de entrenamiento se ha minimizado.

Las redes neuronales *fully connected* han demostrado ser especialmente efectivas para modelar patrones de desempeño al capturar relaciones no lineales complejas entre múltiples factores [30]. Su capacidad para manejar interacciones entre variables, combinada con la flexibilidad de ajustar la arquitectura (número y tamaño de capas), las convierte en herramientas poderosas para predicción y clasificación.

Figura 5. Diagrama iterativo de redes neuronales *fully connected*

3.1.3. Evaluación de modelos de aprendizaje automático supervisado

3.1.3.1. Exactitud, Precisión y Sensibilidad

La exactitud es una métrica utilizada para evaluar el rendimiento de un modelo, indicando el porcentaje de predicciones correctas realizadas sobre un conjunto de datos. Es decir, si se tienen 100 predicciones y el modelo acierta en 60 de ellas, la exactitud del modelo sería del 60%. Aunque es una medida sencilla de interpretar, no siempre es la más confiable y necesita ser complementada con otras métricas [31]. Se calcula de la siguiente manera:

$$exactitud = \frac{\#\text{ predicciones correctas}}{\#\text{ total de predicciones}} = \frac{TP + TN}{TP + TN + FP + FN}$$

Por su parte, la precisión es el porcentaje de predicciones positivas realizadas por un modelo que realmente son correctas y se calculan de la siguiente manera:

$$precision = \frac{\#\text{ positivos predichos correctamente}}{\#\text{ predicciones positivas}} = \frac{TP}{TP + FP}$$

Sensibilidad (Recall): Esta métrica corresponde a la proporción de casos positivos que el modelo clasificó de manera correcta. Se calcula de la siguiente manera:

$$recall = \frac{\#\text{ positivos predichos correctamente}}{\#\text{ casos positivos}} = \frac{TP}{TP + FN}$$

3.1.3.2. Matriz de Confusión

Permite identificar visualmente cuántas predicciones positivas y negativas fueron clasificadas correctamente por el modelo. Esta matriz categoriza las predicciones de la siguiente manera [31]

- (TP) Verdadero Positivo: Son los casos pronosticados como positivos y que realmente son positivos.
- (TN) Verdadero Negativo: Son casos pronosticados como negativos y que realmente son negativos.
- (FP) Falso Positivo: Son casos pronosticados como positivos, pero que realmente fueron negativos, es decir se trata de una predicción incorrecta.
- (FN) Falso Negativo: Son casos pronosticados como negativos pero que realmente eran positivos.

3.1.3.3. Métrica AUC

La métrica AUC (Área Bajo la Curva) es ampliamente utilizada para evaluar el rendimiento de modelos de clasificación, ya que cuantifica el área bajo la curva ROC, la cual muestra la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos en distintos umbrales de decisión [32]. Un valor de AUC próximo a 1 indica que el modelo posee una alta capacidad para distinguir entre clases, mientras que un valor cercano a 0.5 refleja un comportamiento comparable al azar. Una de las ventajas principales de esta métrica es su independencia del umbral de clasificación, lo que la convierte en una herramienta útil en contextos donde los costos asociados a los errores varían o cuando se requiere una evaluación integral del modelo [33]. En la práctica, la AUC se aplica en diversos campos como la medicina, la minería de datos y el aprendizaje automático, facilitando la comparación entre diferentes modelos predictivos. Por ejemplo, en sistemas de diagnóstico médico, una AUC elevada implica una mejor capacidad para identificar correctamente a los pacientes sin incurrir en un número elevado de falsos positivos. Aunque no está exenta de limitaciones como su falta de sensibilidad ante ciertos errores o su incapacidad para reflejar la calidad absoluta del modelo, la AUC sigue siendo una métrica clave para valorar el poder discriminativo de los modelos [34].

3.2. ANTECEDENTES

La predicción de la deserción estudiantil en instituciones de educación superior ha sido abordada por diversas investigaciones a nivel internacional. Un estudio desarrollado por el Politécnico de Milán (Italia) y la Vrije Universiteit Amsterdam (Países Bajos) [35] tuvo como propósito evaluar si los factores que inciden en la deserción se comportan de forma similar en ambos contextos. Para ello, analizaron modelos predictivos aplicados en tres momentos clave del proceso formativo del estudiante: ingreso, final del primer semestre y final del primer año. Entre los algoritmos utilizados se encuentran los modelos lineales generalizados, árboles de clasificación y bosques aleatorios. Los resultados evidenciaron que la capacidad predictiva de estos modelos es en buena medida transferible entre contextos, y que su precisión mejora significativamente al incorporar datos de rendimiento académico. Asimismo, se observó que los modelos entrenados en el contexto neerlandés presentaron mejor desempeño al ser aplicados en el entorno italiano que en sentido inverso.

En el caso latinoamericano, destaca el estudio realizado en una universidad peruana en 2017 [36], cuyo objetivo fue analizar cómo los modelos predictivos aplicados a asignaturas críticas pueden ayudar a identificar a estudiantes en riesgo de deserción. De los modelos evaluados, el que presentó mayor eficacia fue el de árboles de clasificación y regresión. Este se utilizó para predecir los resultados en siete asignaturas, anticipando que 2777 de 4478 estudiantes reprobarían. Gracias a esta predicción, la universidad implementó estrategias de acompañamiento, logrando que al finalizar el semestre 1210 estudiantes aprobaran, reduciendo así el impacto del riesgo inicialmente estimado. Por su parte, en México, una investigación enfocada en estudiantes de ingeniería exploró el uso de algoritmos como regresión logística, árboles de decisión y redes neuronales, concluyendo que es factible identificar con precisión a aquellos estudiantes con alta probabilidad de abandonar sus estudios [37].

Siguiendo esta línea, otro trabajo empleó el algoritmo Classification and Regression Tree (CART) para desarrollar un modelo de predicción de deserción [38]. El modelo resultante tenía una estructura de cuatro niveles de profundidad y generó un conjunto de reglas que permitieron identificar las variables más relevantes en el fenómeno de la deserción dentro de la institución analizada. En el contexto colombiano, Camargo [39] propuso un modelo predictivo para estudiantes de pregrado de la Universidad de la Costa (CUC), considerando factores académicos y socioeconómicos. En su fase de validación, el modelo basado en Random Forest obtuvo un nivel de precisión del 84.8 %, posicionándose como el más efectivo entre las alternativas probadas.

Complementariamente, otro estudio colombiano [40] diseñó un modelo predictivo utilizando árboles de decisión y considerando variables académicas, demográficas y socioeconómicas. Los hallazgos revelaron que aspectos como el promedio académico, el tipo de institución educativa de origen y el estrato socioeconómico inciden significativamente en la probabilidad de deserción. Este modelo permitió identificar con un 80 % de exactitud a los estudiantes con mayor riesgo, lo que facilitó la formulación de intervenciones focalizadas para mejorar la permanencia estudiantil.

Por tanto, los diferentes estudios revisados muestran que los modelos basados en aprendizaje automático, y en particular aquellos fundamentados en árboles de decisión, constituyen herramientas efectivas para la detección temprana de estudiantes en riesgo de deserción. Esta capacidad predictiva permite a las instituciones de educación superior implementar estrategias proactivas orientadas a la reducción del abandono.

4. CAPITULO I. PREPARACIÓN DEL CONJUNTO DE DATOS PARA EL ENTRENAMIENTO Y PRUEBA DE LOS MODELOS PREDICTIVOS

Para dar cumplimiento al primer objetivo específico, orientado a la preparación del conjunto de datos para el entrenamiento y prueba de los modelos mediante procesos de limpieza, transformación y selección de variables socioeconómicas, académicas y personales, se implementaron tres fases metodológicas enmarcadas en el estándar CRISP-DM: comprensión del negocio, comprensión de los datos y preparación de los datos. A continuación, se describen las actividades desarrolladas en cada una de estas fases.

4.1. FASE 1. COMPRENDER EL NEGOCIO

Durante este apartado se realizó la revisión de la literatura relacionada con la predicción de deserción estudiantil en educación superior, abordando sus causas, modelos predictivos y estrategias de intervención institucional. En ese sentido, la deserción estudiantil en la educación superior constituye una problemática crítica a nivel global y particularmente alarmante en América Latina, donde las tasas de abandono pueden alcanzar hasta el 75 % en algunas instituciones [41]. Esta situación afecta de manera directa la eficiencia del sistema educativo y limita el desarrollo social, económico y humano de los países. A pesar de los esfuerzos institucionales por mejorar la permanencia, persisten desafíos significativos en la identificación temprana de factores de riesgo y en la implementación de estrategias efectivas. En este contexto, desarrollar un modelo de predicción de deserción estudiantil fundamentado en técnicas de ciencia de datos y aprendizaje automático resulta altamente útil, ya que permite anticipar comportamientos de abandono, optimizar la toma de decisiones y fomentar intervenciones personalizadas y oportunas [42],[43].

Como fase inicial del proyecto se realizó una exploración en diversos repositorios de bases de datos abiertas, entre ellos Kaggle, UCI Machine Learning Repository y otros portales de open data, con el objetivo de identificar conjuntos de datos pertinentes para la predicción de la deserción estudiantil. Esta búsqueda permitió reconocer las variables comúnmente utilizadas en investigaciones previas sobre permanencia y abandono en educación superior, tales como indicadores académicos, datos sociodemográficos y factores financieros. A partir de dicha exploración, se seleccionó el conjunto de datos "Predict Students' Dropout and Academic Success" disponible en el UCI Machine Learning Repository [44], por ser el que mejor se ajustaba a los objetivos del estudio. Este dataset contiene información histórica de 4.424 estudiantes pertenecientes a programas de agronomía, comunicación, gestión, educación y diseño, entre otros, integrando aspectos académicos (rendimiento y progreso semestral), sociodemográficos (género, estado civil, programa cursado), administrativos y económicos, lo que proporciona una base sólida y contextualizada para la construcción de modelos predictivos efectivos.

4.2. FASE 2. COMPRENDER LOS DATOS

En esta etapa se realizó el análisis exploratorio de los datos con el fin de comprender la estructura del dataset, la distribución de las variables, la presencia de valores faltantes y la calidad general de la información. El dataset utilizado corresponde a una base de datos pública de libre acceso, disponible bajo licencias de distribución abierta para fines académicos, por lo que no requirió autorización específica para su uso.

Es importante precisar que este conjunto de datos corresponde a una institución de educación superior portuguesa, por lo que las variables que lo componen en particular las relacionadas con el sistema de créditos académicos, el esquema de financiamiento estudiantil y los indicadores macroeconómicos responden a la estructura institucional y normativa europea. Dada la limitada disponibilidad de conjuntos de datos institucionales abiertos en el contexto colombiano y latinoamericano con un nivel de granularidad comparable, se optó por utilizar esta base como punto de partida metodológico. Esta decisión implica que los resultados obtenidos constituyen una prueba de concepto cuya transferibilidad a instituciones colombianas debe ser validada empíricamente en estudios futuros, dado que las variables socioeconómicas, los sistemas de financiamiento y los patrones de deserción difieren entre ambos contextos.

Durante la exploración de la estructura del conjunto de datos como se mencionó previamente, está conformado por 4,424 registros y un total de 36 variables. Esta dimensión resulta adecuada para el desarrollo de modelos de predicción, pues permite un volumen de información suficiente para entrenar y validar modelos sin comprometer la capacidad computacional. A partir de los datos se tienen entonces las variables enteras correspondientes principalmente a atributos de naturaleza categórica que han sido codificados numéricamente (por ejemplo, marital status, application mode, nationality); variables decimales que están relacionadas en su mayoría con calificaciones y variables económicas (Admission grade, Curricular units (grade), Unemployment rate, Inflation rate, GDP) y variable objetivo que clasifica los estudiantes en categorías como: Graduate, Dropout y Enrolled.

Tabla 1. Tipos de variables

Variables enteras (int64)	**Variables decimales (float64)**	**Variable Objetivo**
29	7	1

Al revisar el dataset, se identificó que las variables ya contaban con etiquetas adecuadas, por lo que no fue necesario renombrarlas. Posteriormente, se realizó un análisis preliminar que no reportó valores nulos en ninguna columna, lo cual representa una ventaja

significativa para el proceso de preparación de datos, ya que no se requieren procedimientos de imputación. En términos de naturaleza de la información, los datos se estructuran en cuatro dimensiones:

- Académica: desempeño en unidades curriculares de primer y segundo semestre (enrolled, evaluations, approved, grade).
- Socioeconómica: condiciones financieras y de apoyo institucional (Scholarship holder, Debtor, Tuition fees up to date, Educational special needs, Displaced).
- Demográfica: características individuales como Age at enrollment, Gender, Marital status, Nationality.
- Contextual: indicadores macroeconómicos asociados al periodo de ingreso del estudiante (Unemployment rate, Inflation rate, GDP).

Por otro lado, la variable "target" constituye el eje central del estudio, ya que clasifica a los estudiantes según el desenlace de su trayectoria académica en tres categorías: Graduate, Dropout y Enrolled (Tabla 2). Si bien la categoría Enrolled refleja el estado actual de aquellos estudiantes que continúan matriculados en la institución, presenta una limitación metodológica importante, pues no permite conocer con certeza su desenlace futuro; es decir, si culminarán exitosamente su programa académico o si eventualmente desertarán. En consecuencia, la inclusión de esta categoría en un modelo predictivo introduciría ambigüedad y ruido en los datos, comprometiendo la capacidad del algoritmo para discriminar de manera efectiva entre los estados de graduación y deserción. Por esta razón, más adelante se propone la recodificación de esta variable para optimizar el desempeño del modelo.

Tabla 2. Distribución Target

Variable	Cantidad	Proporción
Graduate	2209	49.93%
Dropout	1421	32.12%
Enrolled	794	17.95%
Total	4424	-

Luego, dentro del proceso de exploración de datos, fue necesario evaluar la proporción de datos faltantes o nulos, dado que este análisis constituye un paso fundamental, pues la presencia de registros incompletos puede afectar la calidad de los modelos predictivos. Para ello, se aplicaron funciones de conteo de valores nulos y no nulos en cada variable del conjunto de datos. Los resultados evidenciaron que las 36 variables que conforman el

dataset presentan un 100 % de registros completos, es decir, no se identificaron valores faltantes o nulos en ninguna columna.

Este hallazgo representa una ventaja significativa para la fase de preparación de datos, en la medida en que elimina la necesidad de aplicar técnicas de imputación o eliminación de casos. En consecuencia, se puede afirmar que la calidad de la información en términos de completitud es adecuada, lo que favorece directamente la confiabilidad de los análisis posteriores y garantiza que la variabilidad observada en las predicciones está asociada a las características propias de los estudiantes y no a deficiencias en la captura de datos.

Consecuentemente, se realizó la identificación de las variables relevantes y sus posibles relaciones iniciando con un análisis univariado preliminar, con el fin de conocer la distribución de las variables. Algunos aspectos relevantes identificados fueron:

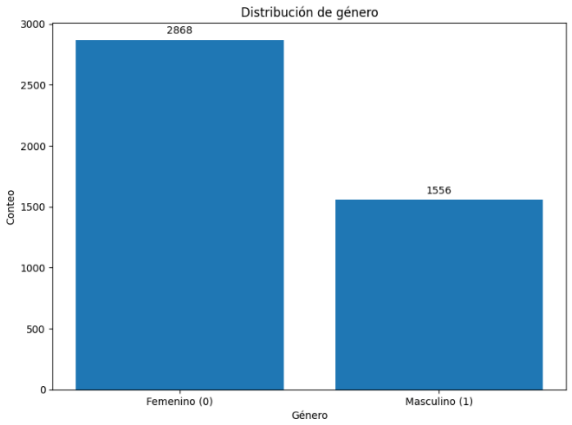


Figura 6. Distribución de género

Se evidencia que hay más estudiantes mujeres (categoría 0) con 2868 registros versus los hombres (categoría 1) con 1556 registros en el conjunto de datos.

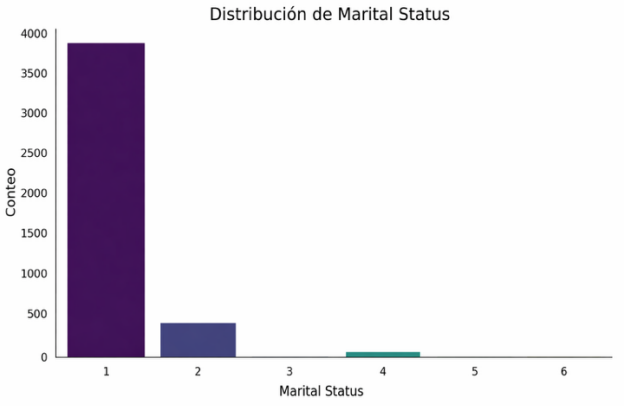


Figura 7. Distribución de estado marital

La mayoría de la población en el conjunto de datos se concentra en la categoría 1 que corresponde a soltero con 3919 datos.

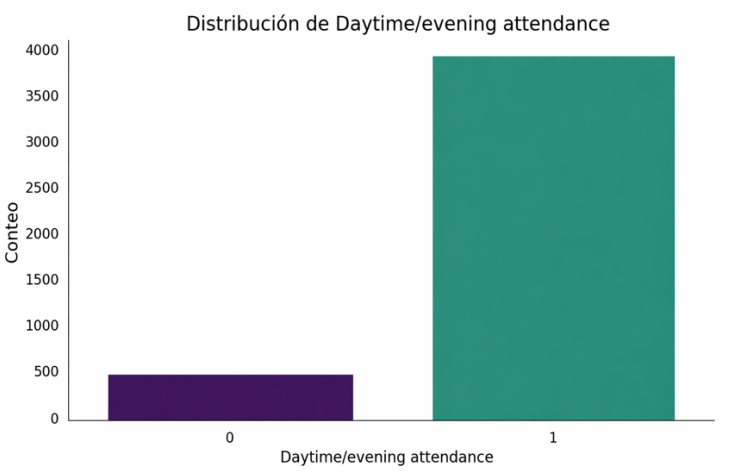


Figura 8. Tipo de asistencia Diurna/Nocturna

El conjunto de datos tiene un número mucho mayor de estudiantes que asisten durante el día (categoría 1) dado por 3941 registros en comparación con la noche (categoría 0) con 483 datos. Sin embargo, con el propósito de identificar la influencia y la relación de cada variable sobre el desenlace, se realizó un análisis bivariado por variable independiente. A partir de este análisis se evidenció que las principales asociaciones se concentraron en los atributos relacionados con las unidades curriculares. No obstante, se observó que dichas variables presentan valores atípicos; tras un proceso de revisión estadística, se consideró que ajustar o eliminar dichos valores podría implicar la pérdida de información relevante, dado que estos podrían representar factores importantes para la explicación de la variable objetivo.

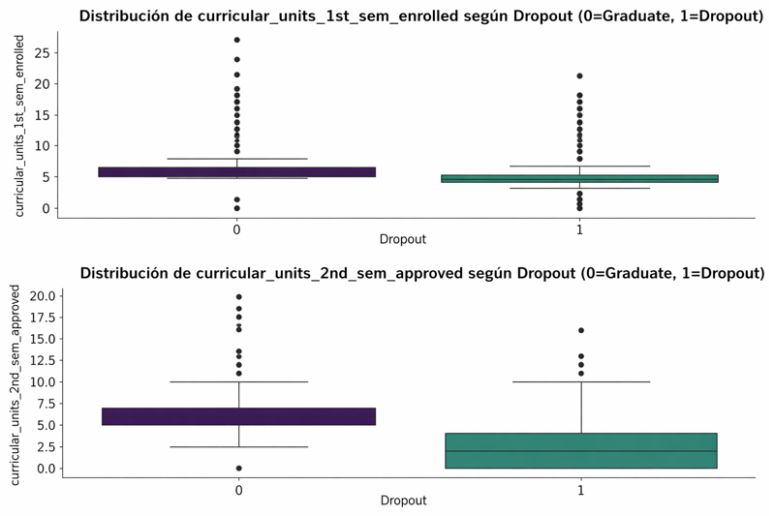


Figura 9. Distribución unidades curriculares

Por tanto, para visualizar de manera conjunta todo el set de datos de variables numéricas se prefirió presentar la matriz de correlación (figura 10).

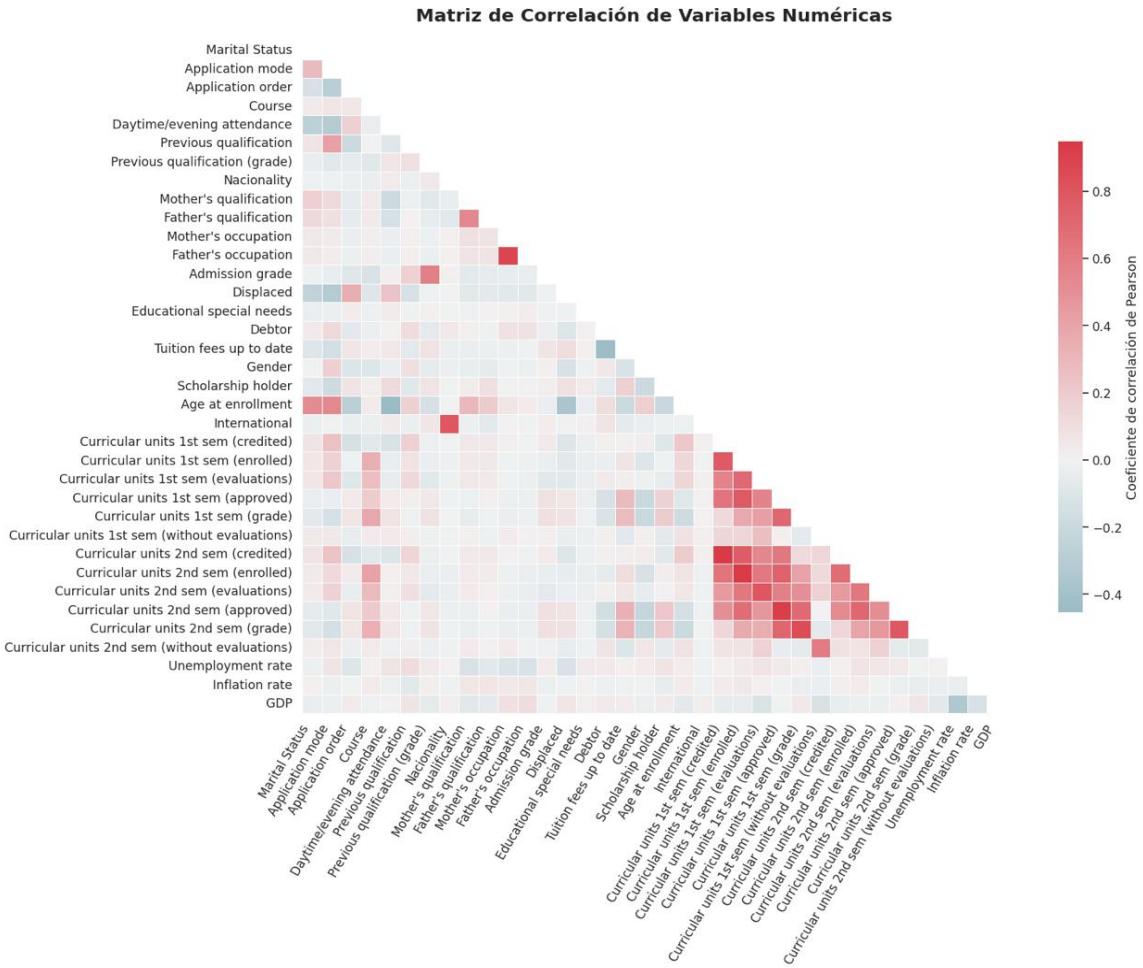


Figura 10. Matriz de correlación

En la matriz de correlación de las variables numéricas del conjunto de datos, se observan correlaciones positivas elevadas entre las variables asociadas al rendimiento académico (unidades curriculares aprobadas, evaluadas y calificadas en los dos semestres), lo que indica una fuerte coherencia interna entre los indicadores de desempeño estudiantil. Por el contrario, las variables sociodemográficas (género, nacionalidad, nivel educativo de los padres) presentan correlaciones débiles o cercanas a cero con las variables de resultado académico, lo que sugiere una limitada relación lineal entre estos factores.

En conjunto, el análisis bivariado evidencia que los estudiantes con menores valores en las variables de rendimiento académico presentan una mayor proporción de deserción, lo que sugiere que el desempeño durante el primer y segundo semestre influye de manera determinante en el resultado final. En contraste, variables como la tasa de desempleo (*unemployment_rate*), el estado civil (*marital_status*) y otros factores sociodemográficos presentan correlaciones débiles con la variable objetivo, por lo que su aporte individual en la predicción podría ser limitado, aunque no se descarta su valor cuando se combinan con otras variables en modelos no lineales.

4.3. FASE 3. PREPARAR LOS DATOS

Durante esta fase se verificó la necesidad de aplicar técnicas de imputación o eliminación de valores nulos o faltantes. Dado que el análisis exploratorio confirmó que el dataset no presentaba valores nulos en ninguna de sus variables, no fue necesario aplicar ninguna técnica de imputación. Esta característica del conjunto de datos constituyó una ventaja metodológica significativa, ya que permitió trabajar directamente con los datos originales sin requerir transformaciones adicionales, garantizando la integridad y autenticidad de la información utilizada en el modelado.

El paso subsecuente, consistió en aplicar técnicas de codificación de variables categóricas, aquí se revisaron las variables del conjunto de datos con el propósito de identificar aquellas de tipo categórico que requirieran un proceso de codificación. Luego del análisis, se evidenció que todas las variables ya se encontraban previamente transformadas a valores numéricos, es decir, no existen columnas de tipo texto ni categorías sin codificar. Variables como Marital status, Application mode, Gender o Nationality están representadas mediante códigos numéricos que permiten su uso directo en los modelos de aprendizaje automático. Por lo tanto, no fue necesario aplicar Label Encoding, dado que el dataset ya está codificado numéricamente en este aspecto.

No obstante, dado que estas variables categóricas están representadas como enteros, los algoritmos de aprendizaje automático podrían interpretarlas erróneamente como variables

continuas con orden implícito. Por esta razón, en la fase de modelado se aplicó one-hot encoding mediante la función pd.get_dummies(), con el parámetro drop_first=True para evitar multicolinealidad. Este proceso transforma cada variable categórica en columnas binarias independientes, garantizando que los modelos las traten como categorías sin relación ordinal entre sí.

Por su parte, como se mencionó previamente la variable "target" tiene tres clases, por tanto, con el propósito de obtener un análisis más claro y orientado a dar respuesta a la pregunta central de la investigación, denotando si ¿el estudiante culminará su programa académico o lo abandonará? se optó por construir un modelo binario. Para ello, se realizó una recodificación de la variable de la siguiente forma:

Tabla 3. Recodificación a variable binaria

Variable	Recodificada	n	Proporción
Graduate	**0**	2209	60.85%
Dropout	**1**	1421	39.15%
Total	-	3630	100%

De esta manera, se excluyó la categoría Enrolled y se conformó un nuevo conjunto de datos con un total de 3,630 registros válidos, distribuidos en 2,209 graduados y 1,421 desertores, lo que equivale al 82.1% de la base original. La decisión de adoptar un enfoque binario responde a la necesidad de centrar el análisis en los dos escenarios definitivos y mutuamente excluyentes: graduación o deserción.

Esta transformación no solo incrementa la claridad interpretativa de los resultados, sino que también proporciona un marco más sólido para el desarrollo de modelos predictivos, garantizando un volumen suficiente de observaciones y un balance adecuado entre clases para el entrenamiento. Con ello, la recodificación de la variable objetivo permite focalizar el estudio en los factores que inciden directamente en la culminación o abandono de los estudios, lo cual resulta fundamental para la generación de estrategias institucionales de intervención y permanencia estudiantil.

Finalmente, en esta fase se aplicaron técnicas de detección y tratamiento de valores atípicos. Para ello se emplearon métodos estadísticos como el análisis de los rangos intercuartílicos (IQR) y el z-score estandarizado, los cuales permitieron detectar observaciones que se alejaban significativamente de la tendencia central. En la mayoría de los casos, los valores atípicos correspondían a situaciones reales de la población estudiantil (por ejemplo, edades de ingreso muy elevadas o calificaciones altas), por lo que se optó por

conservarlos, dado que representaban condiciones particulares y plausibles de los estudiantes.

En cuanto al tratamiento del desbalance de clases evidenciado en la variable objetivo (60.85% Graduate vs. 39.15% Dropout, Tabla 3), se optó por una estrategia de ajuste de pesos (*class weighting*) en lugar de técnicas de sobremuestreo sintético como SMOTE (*Synthetic Minority Oversampling Technique*) o ADASYN (*Adaptive Synthetic Sampling*). Esta decisión se sustenta en tres consideraciones técnicas. Primero, el grado de desbalance observado (razón aproximada 1.5:1) se considera moderado y no severo, rango en el cual la literatura sobre aprendizaje con clases desbalanceadas reporta que el ajuste de pesos suele ofrecer un desempeño comparable al de las técnicas de remuestreo, sin los riesgos adicionales que estas últimas conllevan. Segundo, dado que un subconjunto importante de las variables predictoras (unidades curriculares aprobadas, evaluadas, matriculadas y calificadas) presenta correlaciones internas elevadas (ver Figura 10), la generación de observaciones sintéticas mediante interpolación en el espacio de características —el mecanismo central de SMOTE y ADASYN— podría producir combinaciones irreales entre estas variables, distorsionando la estructura de covarianza original del conjunto de datos. Tercero, el ajuste de pesos modifica directamente la función de pérdida durante el entrenamiento (mediante *class_weight = "balanced"* en regresión logística y *scale_pos_weight* en XGBoost), preservando la distribución original de los datos y evitando la necesidad de aplicar el remuestreo exclusivamente dentro de cada fold de validación cruzada, lo cual habría incrementado sustancialmente la complejidad computacional del esquema de validación cruzada repetida (15 particiones) y de la búsqueda en rejilla de hiperparámetros ya implementado en este estudio.

Es importante señalar que, durante la fase de preparación de los datos, no se llevó a cabo un proceso exhaustivo de ingeniería de características orientado a la creación de nuevas variables derivadas que pudieran capturar de forma más precisa la realidad académica y contextual de los estudiantes. En particular, no se exploraron indicadores relativos tales como la razón entre la edad del estudiante y la edad promedio de su cohorte o programa académico, ni la proporción de unidades curriculares aprobadas respecto a las matriculadas en cada semestre, variables que, de acuerdo con la literatura sobre deserción estudiantil, podrían aportar señales adicionales sobre el ajuste del estudiante a su entorno académico y su carga real de trabajo. Esta omisión constituye una limitación del presente estudio, la cual se retoma como recomendación explícita en la sección de trabajos futuros.

5. CAPITULO II. IMPLEMENTAR ALGORITMOS DE APRENDIZAJE AUTOMATICO SUPERVISADO Y EVALUAR EL RENDIMIENTO DE LOS MODELOS A TRAVÉS DE MÉTRICAS DE DESEMPEÑO

Para la evaluación del desempeño se adoptó un esquema metodológico común a los cinco modelos entrenados, con el fin de garantizar la comparabilidad directa de sus métricas. En primer lugar, el conjunto de datos depurado fue dividido de forma estratificada en un 80% para entrenamiento (2.904 registros) y un 20% para prueba independiente (726 registros), preservando en ambas particiones la proporción de desertores y no desertores. La partición de prueba (*holdout*) se mantuvo aislada durante todo el proceso de ajuste y se utilizó exclusivamente para la evaluación final de generalización de cada modelo.

Sobre el conjunto de entrenamiento se aplicó un esquema de validación cruzada estratificada repetida (RepeatedStratifiedKFold, 5 folds × 3 repeticiones), lo que permitió estimar el desempeño promedio a partir de 15 particiones independientes y reducir la variabilidad asociada a una única división de los datos. Este mismo esquema se utilizó para la optimización de hiperparámetros mediante GridSearchCV en los cuatro modelos basados en scikit-learn (regresión logística, árbol de decisión, Random Forest y XGBoost), asegurando que la selección de configuraciones óptimas se realizara sin contaminación por parte del conjunto de prueba y evitando así cualquier fuga de información (*data leakage*). En el caso de la regresión logística, dado que el escalado de variables es una condición necesaria para la correcta estimación de sus coeficientes, el StandardScaler se integró dentro de un Pipeline junto con el clasificador, de modo que el ajuste de la estandarización se realizara de forma independiente en cada fold de la validación cruzada, preservando así la integridad del esquema.

En el caso de la red neuronal *fully connected*, el esquema de validación cruzada repetida empleado en los demás modelos resulta computacionalmente poco práctico, dado el elevado costo de entrenar iterativamente la red en cada uno de los 15 folds por combinación de hiperparámetros evaluada. Adicionalmente, las arquitecturas de aprendizaje profundo disponen de mecanismos propios para el control del sobreajuste que no requieren validación cruzada externa: en este estudio se empleó regularización L2, *dropout* y *early stopping* sobre un subconjunto de validación interno (validation_split = 0.20 dentro del conjunto de entrenamiento), tal como se detalla en su sección correspondiente. No obstante, la partición 80/20 inicial y la evaluación final sobre el conjunto de prueba se mantuvieron idénticas a las de los otros cuatro modelos.

Como criterio de optimización de hiperparámetros se utilizó el F1-score, métrica que equilibra precisión y recall en contextos con desbalance entre clases. Para la selección del modelo final, en cambio, se priorizó el Recall como criterio determinante, dado que en el

contexto de la deserción estudiantil el costo institucional de los falsos negativos (estudiantes en riesgo que no son identificados) resulta significativamente mayor que el de los falsos positivos, lo que justifica la preferencia por modelos con mayor sensibilidad para la detección temprana de casos vulnerables.

Con el fin de garantizar la reproducibilidad de los experimentos, se fijaron semillas de aleatoriedad previo al entrenamiento de cada modelo. Para el árbol de decisión, Random Forest y XGBoost se utilizó random_state = 42 de forma consistente en la partición de datos, los estimadores y el esquema de validación cruzada. En el caso de la regresión logística, se realizó un análisis de estabilidad adicional entrenando el modelo sobre 31 particiones estratificadas 80/20 con distintas semillas (random_state ∈ {1, 2, ..., 30, 42}), manteniendo fijos los hiperparámetros óptimos (C=0.1, penalty=L2, solver=lbfgs). El análisis reveló que la partición generada con random_state = 42 producía un comportamiento atípico, con un Recall en el conjunto de prueba (0.9296) superior al observado en entrenamiento (0.8830), patrón inconsistente con el comportamiento promedio observado en el resto de las semillas (Recall_test medio = 0.8604, desviación estándar = 0.0206). La brecha Train−Test asociada a esta semilla (−0.0466) se ubica a 3.36 desviaciones estándar de la media de la distribución empírica, configurando un outlier estadístico. Se interpretó este resultado como evidencia de que la partición aleatoria específica asociada a random_state = 42 generaba un conjunto de prueba no representativo del comportamiento promedio esperado bajo muestreo aleatorio. En consecuencia, para la regresión logística se adoptó random_state = 12 en la partición de datos, cuya brecha Train−Test (+0.0283) se ubica dentro del rango típico de la distribución (Z = −0.14 respecto a la media de la brecha Train−Test) y proporciona una estimación más representativa de la capacidad de generalización del modelo. El estimador y el esquema de validación cruzada mantuvieron random_state = 42. Para la red neuronal, se fijaron np.random.seed (42) en NumPy y tf.random.set_seed (42) en TensorFlow previo al entrenamiento, con el fin de reducir la variabilidad asociada a la inicialización aleatoria de pesos.

Con el propósito de favorecer la reproducibilidad del estudio, se documentan las versiones específicas de las principales librerías empleadas durante el proceso de modelado: *scikit-learn* 1.6.1, *XGBoost* 3.2.0, *TensorFlow* 2.20.0, *pandas* 2.2.2 y *NumPy* 2.0.2. Esta documentación permite a futuros investigadores replicar las condiciones exactas de entrenamiento, dado que diferencias entre versiones de estas librerías pueden ocasionar variaciones menores en los resultados, incluso manteniendo fijas las semillas de aleatoriedad.

En relación con la posible influencia de la partición específica del conjunto de prueba sobre el desempeño relativo de cada modelo, se identifican varios elementos del diseño

experimental que mitigan este riesgo. En primer lugar, la estratificación aplicada en la partición garantiza que la proporción de clases (60.85% Graduate / 39.15% Dropout) se mantenga constante tanto en entrenamiento como en prueba, evitando que diferencias en el balance de clases entre particiones favorezcan sistemáticamente a un modelo en particular. En segundo lugar, la optimización de hiperparámetros de los modelos basados en *scikit-learn* no se realizó sobre una única partición, sino mediante validación cruzada estratificada repetida (15 particiones independientes), lo que reduce la dependencia de los hiperparámetros seleccionados respecto a las particularidades de una única división de los datos. En tercer lugar, el análisis de estabilidad realizado específicamente para la regresión logística (descrito anteriormente en esta misma sección, sobre 31 semillas distintas) constituye evidencia empírica directa de que el conjunto de prueba asociado a una semilla particular puede comportarse de manera atípica para un modelo dado, lo cual motivó precisamente la adopción de una semilla alternativa para ese caso específico. No obstante, se reconoce que un análisis de sensibilidad similar sobre los modelos restantes, evaluando su desempeño bajo múltiples particiones aleatorias del conjunto de prueba, constituiría una validación adicional deseable para descartar de manera más concluyente cualquier efecto de partición específica sobre la comparación entre modelos.

A continuación, se presentan los resultados obtenidos para cada uno de los cinco modelos evaluados.

5.1. REGRESIÓN LOGÍSTICA

Este modelo fue seleccionado por su solidez teórica, interpretabilidad y eficacia en problemas de clasificación binaria. En este estudio, la variable dependiente corresponde a la condición de deserción estudiantil, definida como una variable binaria (1 = desertor, 0 = no desertor).

5.1.1. Preparación de los datos

Para el entrenamiento del modelo se utilizó la base de datos previamente depurada. La matriz de características X se construyó excluyendo la variable objetivo (Target), mientras que la variable respuesta y fue convertida a tipo entero para garantizar su correcta interpretación binaria. Como se describió en el Capítulo I, se aplicó one-hot encoding mediante pd.get_dummies() sobre las variables categóricas codificadas numéricamente, generando columnas binarias independientes para garantizar su correcta interpretación por parte de los algoritmos. Adicionalmente, se realizó un proceso de estandarización de variables (StandardScaler), garantizando que todas las características presentaran media cero y desviación estándar unitaria. Este escalado se efectuó dentro de cada partición de validación cruzada para evitar fuga de información (*data leakage*).

El conjunto de datos fue dividido de forma estratificada en 80% para entrenamiento (2.904 registros) y 20% para prueba (726 registros), preservando la proporción de desertores y no desertores en ambas particiones.

5.1.2. Optimización de hiperparámetros

Con el propósito de garantizar estabilidad y robustez en la estimación del desempeño, se implementó un esquema de validación cruzada estratificada repetida (RepeatedStratifiedKFold) configurado con 5 folds y 3 repeticiones, para un total de 15 evaluaciones por combinación de hiperparámetros. Para la optimización se utilizó GridSearchCV con el criterio principal F1-score, evaluando 16 combinaciones para un total de 240 ajustes. Se configuró además el parámetro class_weight = "balanced" para compensar automáticamente el desbalance en la variable objetivo. La tabla siguiente presenta los valores explorados y el óptimo seleccionado para cada hiperparámetro.

Tabla 4. Búsqueda y selección de hiperparámetros (RL)

Hiperparámetro	Valores evaluados	Valor óptimo
Penalty	L1, L2	L2
Solver	lbfgs, liblinear, saga	lbfgs
C	0.01, 0.1, 1, 10	0.1

El mejor F1-score promedio en validación cruzada fue F1 = 0.8815, con una desviación estándar de 0.0142, lo cual indica estabilidad en el desempeño del modelo. Cabe destacar que la regularización L2 (Ridge) penaliza los coeficientes grandes sin llevarlos a cero, favoreciendo soluciones más suaves y contribuyendo a la generalización del modelo al reducir el riesgo de sobreajuste.

5.1.3. Evaluación de desempeño y generalización

Con el propósito de verificar la capacidad de generalización del modelo, se evaluó su desempeño durante la fase de optimización mediante validación cruzada estratificada repetida, y posteriormente sobre los conjuntos de entrenamiento y prueba. Las métricas obtenidas se presentan en las tablas siguientes.

Tabla 5. Métricas promedio en validación cruzada (RL)

Métrica	Media CV (5 folds x 3 repeticiones)	Desv. Estándar
Accuracy	0.9083	± 0.0096
Precision	0.8922	± 0.0195
Recall	0.8716	± 0.0194
F1-score	0.8815	± 0.0123
ROC-AUC	0.9516	± 0.0078

Tabla 6. Métricas de desempeño (RL)

Métrica	Entrenamiento	Prueba	Diferencia absoluta
Accuracy	0.9136	0.9036	0.0100
Precision	0.8977	0.8905	0.0072
Recall	0.8795	0.8592	0.0203
F1-score	0.8885	0.8746	0.0139
ROC-AUC	0.9570	0.9508	0.0062

Los resultados de la validación cruzada (Tabla 5) evidencian un desempeño estable durante la fase de optimización, con desviaciones estándar bajas en todas las métricas (≤ 0.0195), confirmando que la selección de hiperparámetros no dependió de una partición particular de los datos. La brecha absoluta entre entrenamiento y prueba es baja en todas las métricas (≤ 0.021), lo que indica ausencia de sobreajuste significativo. En particular, la diferencia en ROC-AUC es de 0.0062, confirmando que el modelo mantiene su capacidad discriminativa al enfrentarse a datos no vistos durante el entrenamiento. Este comportamiento es coherente con el efecto de la regularización L2 con C = 0.1, la cual penaliza coeficientes grandes y favorece la generalización del modelo.

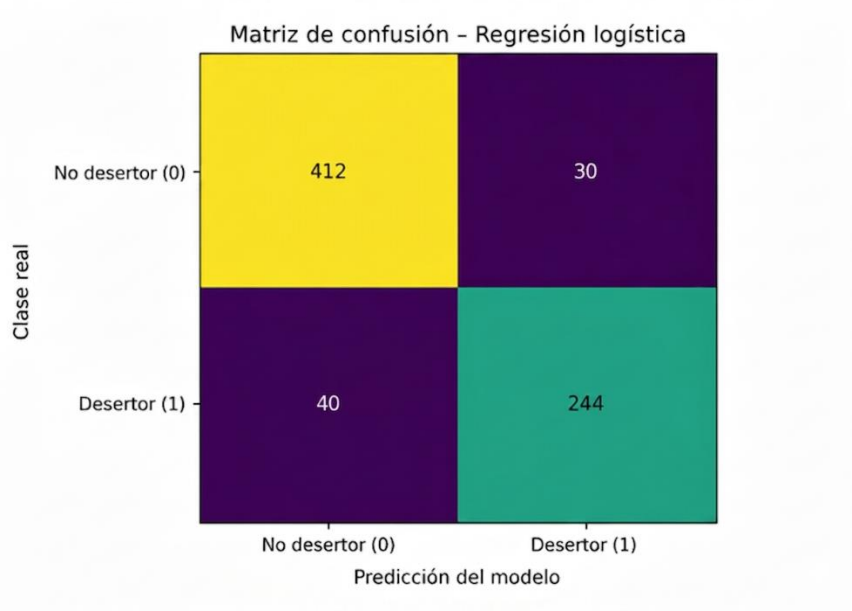


Figura 11. Matriz de confusión Test (RL)

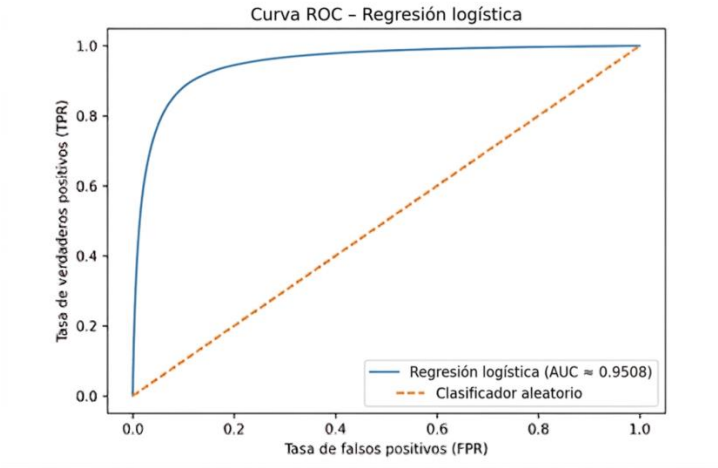


Figura 12. Curva ROC Test (RL)

La figura 11 presenta la matriz de confusión para el conjunto de prueba. El modelo clasifica correctamente 412 estudiantes no desertores (verdaderos negativos) y 244 desertores (verdaderos positivos), registrando 30 falsos positivos y 40 falsos negativos. Lo anterior representa una sensibilidad (recall clase 1) del 85.92% y una especificidad (clase 0) del 93.21%.

Desde una perspectiva institucional, el modelo presenta un equilibrio adecuado entre la reducción de falsos negativos que implican no intervenir a tiempo y el control de falsos positivos que pueden generar intervenciones innecesarias. Si bien el costo institucional de un falso negativo puede ser mayor, el nivel observado se encuentra dentro de rangos aceptables para un modelo predictivo aplicado al contexto educativo.

La figura 12 presenta la curva ROC obtenida para el modelo de regresión logística. El área bajo la curva (ROC-AUC = 0.9508 en el conjunto de prueba) evidencia una elevada capacidad discriminativa del modelo, indicando que existe una alta probabilidad de que el algoritmo asigne una puntuación de riesgo mayor a un estudiante desertor que a uno que no abandona sus estudios.

5.1.4. Análisis de importancia de las variables

A diferencia de los modelos basados en árboles, la regresión logística no produce métricas de importancia basadas en reducción de impureza. Sin embargo, dado que las variables fueron estandarizadas previamente (media cero, desviación estándar unitaria), los coeficientes del modelo son directamente comparables entre sí y constituyen una medida válida de la influencia relativa de cada variable sobre la probabilidad de deserción.

Un coeficiente negativo indica que el incremento de la variable reduce la probabilidad de deserción, mientras que un coeficiente positivo señala el efecto contrario. La tabla siguiente presenta las diez variables con mayor coeficiente en valor absoluto.

Tabla 7. Top 10 variables más influyentes (RL)

Variable	Coeficiente	Dirección del efecto	Importancia relativa
Curricular units 2nd sem (approved)	-2.1261	- Reduce riesgo de deserción	27.9%
Curricular units 1st sem (approved)	-1.3394	- Reduce riesgo de deserción	17.6%
Curricular units 2nd sem (enrolled)	0.8660	+ Aumenta riesgo de deserción	11.4%
Tuition fees up to date	-0.8026	- Reduce riesgo de deserción	10.5%
Curricular units 2nd sem (grade)	-0.5581	- Reduce riesgo de deserción	7.3%
Curricular units 1st sem (enrolled)	0.4417	+ Aumenta riesgo de deserción	5.8%
Course	0.4048	+ Aumenta riesgo de deserción	5.3%
Debtor	0.3829	+ Aumenta riesgo de deserción	5.0%
Scholarship holder	-0.3561	- Reduce riesgo de deserción	4.7%
Curricular units 1st sem (credited)	0.3378	+ Aumenta riesgo de deserción	4.4%

La variable con mayor influencia es Curricular units 2nd sem (approved) (coeficiente = -2.1261), lo que indica que un mayor número de materias aprobadas en el segundo semestre reduce significativamente la probabilidad de deserción. En contraste, Curricular units 2nd sem (enrolled) presenta un coeficiente positivo (0.8660), sugiriendo que una alta carga académica sin el correspondiente rendimiento se asocia con mayor riesgo de abandono.

En tercer lugar, Curricular units 1st sem (approved) refuerza el hallazgo anterior: el desempeño académico temprano (ya desde el primer semestre) es un predictor consistente del riesgo de deserción. La variable Tuition fees up to date (coeficiente = -0.8026) confirma que estar al día en el pago de matrícula reduce considerablemente la probabilidad de abandono, incorporando una dimensión económica al fenómeno. Este comportamiento refleja la capacidad del modelo para capturar efectos simultáneos y combinados, integrando dimensiones académicas (desempeño en créditos aprobados, carga inscrita), económicas (estado de pago, condición de deudor) y socioeconómicas (becario, ocupación de la madre). Este patrón multifactorial es coherente con la literatura sobre deserción estudiantil, que reconoce el carácter multicausal del fenómeno.

El modelo de Regresión Logística implementado demuestra una alta capacidad predictiva, estabilidad estadística, un adecuado equilibrio entre precisión y recall, y un excelente poder discriminativo. La regularización L2 con C = 0.1 favorece la generalización del modelo, reduciendo el riesgo de sobreajuste y permitiendo un desempeño consistente entre entrenamiento y prueba, con brechas absolutas inferiores a 0.021 en todas las métricas evaluadas. El análisis de coeficientes revela que el fenómeno de deserción estudiantil está explicado de manera multifactorial, con el desempeño académico temprano como principal determinante, complementado por factores económicos y administrativos. Por tanto, la Regresión Logística constituye una herramienta viable para apoyar estrategias institucionales de prevención de la deserción estudiantil, permitiendo identificar de manera

temprana a estudiantes con mayor probabilidad de abandono y facilitando la focalización de intervenciones académicas y administrativas.

5.2. ARBOLES DE DECISIÓN

Con el propósito de complementar el análisis predictivo y explorar modelos no lineales, se implementó un modelo de Árboles de Decisión (AD) para clasificación binaria, técnica basada en particiones recursivas del espacio de características. A diferencia de la regresión logística, modelo paramétrico lineal, el árbol permite capturar interacciones complejas y relaciones no lineales entre variables sin asumir una forma funcional específica, lo que resulta particularmente relevante en el estudio de la deserción estudiantil, donde la interacción entre variables académicas, económicas y administrativas puede no responder a patrones estrictamente lineales.

5.2.1. Preparación de los datos

Para el entrenamiento del modelo se utilizó la base de datos previamente preparada. Se aplicó one-hot encoding mediante pd.get_dummies() sobre las variables categóricas, tal como se describió en el Capítulo I. El conjunto de datos fue dividido de forma estratificada en 80% para entrenamiento (2.904 registros) y 20% para prueba (726 registros), preservando la proporción de desertores y no desertores en ambas particiones.

5.2.2. Optimización de hiperparámetros

La selección de hiperparámetros se realizó mediante GridSearchCV con validación cruzada estratificada repetida (5 folds × 3 repeticiones), para un total de 15 evaluaciones por combinación, empleando como criterio principal el F1-score. Se evaluaron 120 combinaciones para un total de 1.800 ajustes. La tabla siguiente presenta los valores explorados y el óptimo seleccionado para cada hiperparámetro.

Tabla 8. Búsqueda y selección de hiperparámetros (AD)

Hiperparámetro	Valores evaluados	Valor óptimo
Criterion	gini, entropy	gini
max_depth	4, 6, 8, 10, None	6
min_samples_split	10, 20, 40	10
min_samples_leaf	4, 8, 16, 24	24

El mejor F1 score promedio en validación cruzada fue: $F1_{CV} = 0.8415$. Este valor es inferior al obtenido con el modelo de regresión logística (0.8815), lo cual sugiere que la estructura lineal regularizada captura con mayor eficiencia la frontera de decisión bajo las condiciones del conjunto de datos analizado. Es importante destacar que el modelo óptimo seleccionó una profundidad moderada del árbol (max_depth = 6), lo cual indica una fuerte regularización estructural. Esto sugiere que árboles más profundos tendían a sobreajustar

los datos, mientras que estructuras más compactas presentaron mejor capacidad de generalización.

5.2.3. Evaluación del desempeño y generalización

Una vez seleccionado el modelo óptimo, se evaluó su desempeño durante la fase de optimización mediante validación cruzada estratificada repetida, y posteriormente sobre los conjuntos de entrenamiento y prueba. Las métricas obtenidas se presentan en las tablas siguientes.

Tabla 9. Métricas promedio en validación cruzada AD

Métrica	Media CV (5 folds x 3 repeticiones)	Desv. Estándar
Accuracy	0.8784	± 0.0089
Precision	0.8601	± 0.0230
Recall	0.8250	± 0.0287
F1-score	0.8415	± 0.0121
ROC-AUC	0.9247	± 0.0141

Tabla 10. Métricas de desempeño del modelo AD

Métrica	Entrenamiento	Prueba	Diferencia absoluta
Accuracy	0.8950	0.9022	0.0072
Precision	0.8895	0.8790	0.0105
Recall	0.8355	0.8697	0.0342
F1-score	0.8617	0.8743	0.0126
ROC-AUC	0.9451	0.9494	0.0043

Los resultados de la validación cruzada mostrados en la Tabla 9 evidencian un desempeño estable durante la fase de optimización, con desviaciones estándar bajas en todas las métricas (≤ 0.0287), confirmando que la selección de hiperparámetros no dependió de una partición particular de los datos. Teniendo en cuenta lo anterior, observamos que la brecha absoluta entre entrenamiento y prueba fue baja en todas las métricas (≤ 0.035), lo que indica ausencia de sobreajuste significativo. Es destacable que el Recall y el ROC-AUC sean ligeramente superiores en el conjunto de prueba que en entrenamiento, lo cual puede interpretarse como efecto de la regularización estructural fuerte derivada de la profundidad limitada del árbol (max_depth = 6), la estabilidad en la distribución de clases entre particiones y la ausencia de memorización del conjunto de entrenamiento. Desde el punto de vista estadístico, el modelo presenta buena capacidad de generalización, aunque con menor desempeño global comparado con la regresión logística.

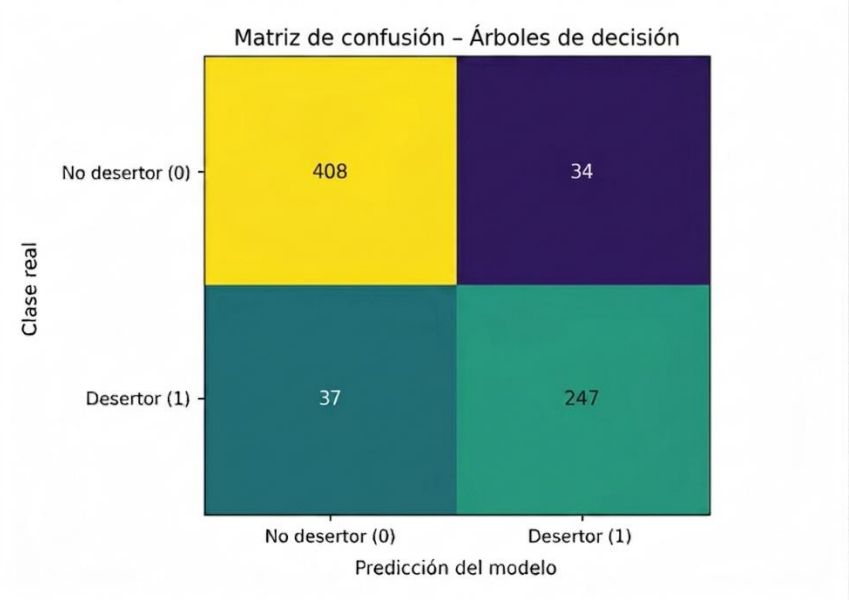


Figura 13. Matriz de confusión (AD)

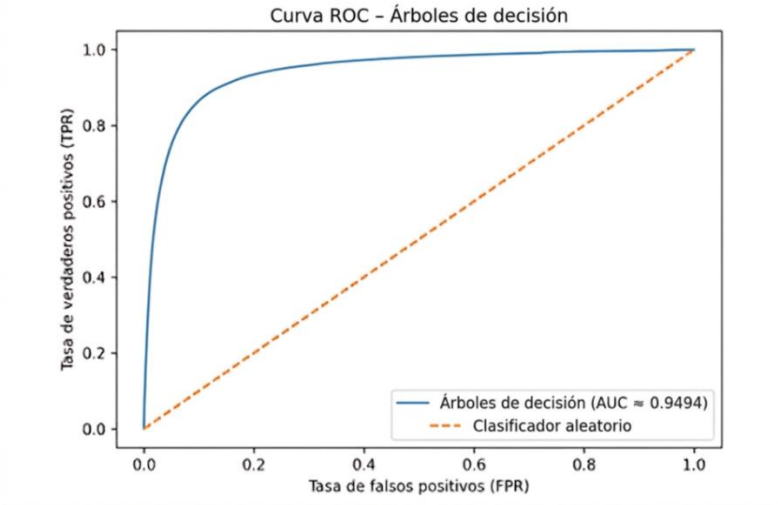


Figura 14. Curva ROC (AD)

La figura 13 demuestra que, para el modelo de árbol de decisión en el conjunto de test, los datos están distribuidos como verdaderos negativos (408), falsos positivos (34), falsos negativos (37) y verdaderos positivos (247). Lo cual representa una Sensibilidad (Recall clase 1): 86.97% y una especificidad (clase 0): 92.31%. Por tanto, el modelo logra un equilibrio razonable entre ambas clases, aunque el costo institucional de los falsos negativos continúa siendo un punto crítico de análisis estratégico.

La figura 14 correspondiente a ROC-AUC en test fue 0.9494, este resultado indica una alta capacidad discriminativa del modelo de árbol de decisión para distinguir entre estudiantes

desertores y no desertores. No obstante, dicho valor es inferior al obtenido por la regresión logística (0.9508), lo que sugiere que, para el conjunto de datos analizado, la regresión logística presenta una capacidad ligeramente superior para separar ambas clases. Aun así, la diferencia entre ambos modelos es moderada y confirma que el árbol de decisión constituye una alternativa predictiva con buen desempeño.

5.2.4. Análisis de importancia de variables

El análisis de importancia basado en reducción de impureza Gini mostró una concentración significativa del poder predictivo en una sola variable: Curricular units 2nd sem (approved) (0.8053). Este nivel de concentración (más del 80% de importancia) indica que el modelo depende fuertemente del desempeño académico en segundo semestre para segmentar la población estudiantil.

Tabla 11. Top 10 variables más influyentes (AD)

Variable	Importancia (Gini)
Curricular units 2nd sem (approved)	80.53%
Tuition fees up to date	7.83%
Curricular units 2nd sem (enrolled)	2.62%
Curricular units 2nd sem (grade)	1.91%
Curricular units 2nd sem (evaluations)	1.90%
Curricular units 1st sem (evaluations)	1.30%
Course	1.21%
Curricular units 1st sem (approved)	1.18%
Previous qualification (grade)	0.59%
Debtor	0.35%

Desde un punto de vista metodológico, esta concentración puede interpretarse de dos maneras: la variable realmente posee alto poder explicativo, o el árbol prioriza variables con mayor capacidad de partición temprana, lo que puede subestimar la contribución combinada de otras variables. En contraste, la regresión logística distribuye los coeficientes de forma más homogénea, capturando efectos combinados de múltiples factores. Este hallazgo sugiere que el fenómeno de deserción, aunque altamente influenciado por el desempeño académico, podría estar mejor explicado por modelos que integren efectos simultáneos y no exclusivamente jerárquicos.

5.3. BOSQUES ALEATORIOS

Con el fin de evaluar un modelo de mayor complejidad estructural y capacidad predictiva, se implementó un Random Forest Classifier, técnica de aprendizaje supervisado basada en el principio de *bagging* (*Bootstrap Aggregating*). A diferencia del árbol de decisión simple el cual puede presentar alta sensibilidad a pequeñas variaciones en los datos, el bosque aleatorio (BA), introduce aleatorización en la selección de muestras, aleatorización en la selección de características en cada división y promediado de múltiples estimadores. Estas

propiedades lo convierten en un modelo adecuado para fenómenos complejos como la deserción estudiantil, donde existen múltiples interacciones entre variables académicas, económicas y contextuales.

5.3.1. Preparación de los datos

Se utilizó la misma base de datos previamente depurada, con one-hot encoding aplicado mediante pd.get_dummies() conforme a lo descrito en el Capítulo I y el conjunto fue dividido en 80% entrenamiento y 20% prueba. La partición fue estratificada para preservar la proporción de desertores y no desertores.

5.3.2. Optimización de hiperparámetros

Se evaluaron 60 combinaciones de hiperparámetros mediante GridSearchCV con validación cruzada estratificada repetida (5 folds × 3 repeticiones), para un total de 900 ajustes, empleando como criterio principal el F1 score para equilibrar precisión y sensibilidad en contextos con desbalance entre clases. La tabla siguiente presenta los valores explorados y el óptimo seleccionado para cada hiperparámetro.

Tabla 12. Búsqueda y selección de hiperparámetros (BA)

Hiperparámetro	Valores evaluados	Valor óptimo
n_estimators	200, 350, 500	350
Criterion	gini, entropy	gini
max_depth	None, 12, 16, 20, 24	12
min_samples_split	20	20
min_samples_leaf	4, 12	4
max_features	sqrt	sqrt

El mejor F1-score promedio en validación cruzada fue F1CV = 0.8661. Este resultado supera al árbol de decisión simple (0.8415) y se aproxima al desempeño de la regresión logística (0.8815), evidenciando la ventaja del ensamble sobre modelos individuales. La selección de criterion = gini indica que el índice de impureza resultó más eficiente para particionar el espacio de características en este conjunto de datos. Asimismo, la profundidad máxima de 12 permite capturar interacciones complejas sin incurrir en sobreajuste excesivo, gracias al mecanismo de promediado propio del bosque aleatorio.

5.3.3. Evaluación del desempeño y generalización

El desempeño del modelo se evaluó durante la fase de optimización mediante validación cruzada estratificada repetida, y posteriormente sobre los conjuntos de entrenamiento y prueba. Las métricas obtenidas se presentan en las tablas siguientes.

Tabla 13. Métricas promedio en validación cruzada (BA)

Métrica	Media CV (5 folds x 3 repeticiones)	Desv. Estándar
Accuracy	0.8981	± 0.0127
Precision	0.8919	± 0.0202
Recall	0.8420	± 0.0192
F1-score	0.8661	± 0.0168
ROC-AUC	0.9464	± 0.0093

Tabla 14. Métricas de desempeño (BA)

Métrica	Entrenamiento	Prueba	Diferencia absoluta
Accuracy	0.9318	0.9132	0.0186
Precision	0.9384	0.8850	0.0534
Recall	0.8839	0.8944	0.0105
F1-score	0.9103	0.8897	0.0206
ROC-AUC	0.9831	0.9687	0.0144

Los resultados de la validación cruzada (Tabla 13) evidencian un desempeño estable durante la fase de optimización, con desviaciones estándar bajas en todas las métricas (≤ 0.0202), confirmando que la selección de hiperparámetros no dependió de una partición particular de los datos. Aunque existe una diferencia algo mayor en precisión (0.0534), todas las brechas entre entrenamiento y prueba se mantienen dentro de rangos aceptables para modelos de aprendizaje automático, lo que indica ausencia de sobreajuste significativo. El valor elevado de ROC-AUC en entrenamiento (0.9831) evidencia una capacidad discriminativa muy alta dentro del conjunto de entrenamiento, mientras que el descenso moderado en el conjunto de prueba (0.9687) es consistente con el comportamiento esperado de modelos de mayor complejidad. Esto sugiere que el modelo logra un equilibrio adecuado entre ajuste y generalización.

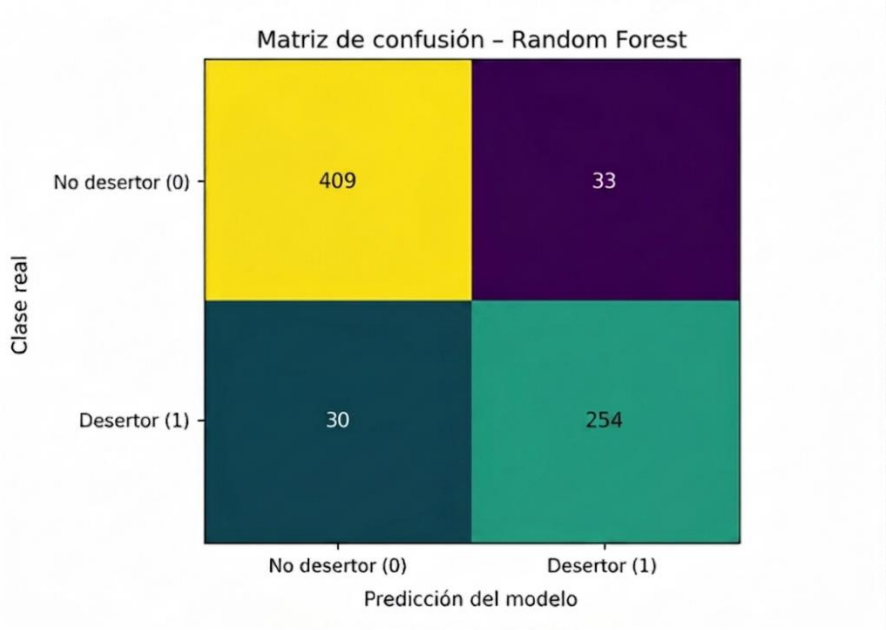


Figura 15. Matriz de confusión (Test) (BA)

El modelo clasifica correctamente 409 estudiantes no desertores y 254 desertores, registrando 33 falsos positivos y 30 falsos negativos.

En comparación con el árbol de decisión simple, el modelo de bosque aleatorio reduce simultáneamente falsos positivos y falsos negativos, lo que evidencia un mejor equilibrio entre sensibilidad y especificidad. Así mismo, el recall del 89.44% confirma una elevada capacidad para detectar estudiantes en riesgo de deserción, aspecto especialmente relevante en contextos institucionales donde la identificación temprana permite implementar estrategias de intervención académica.

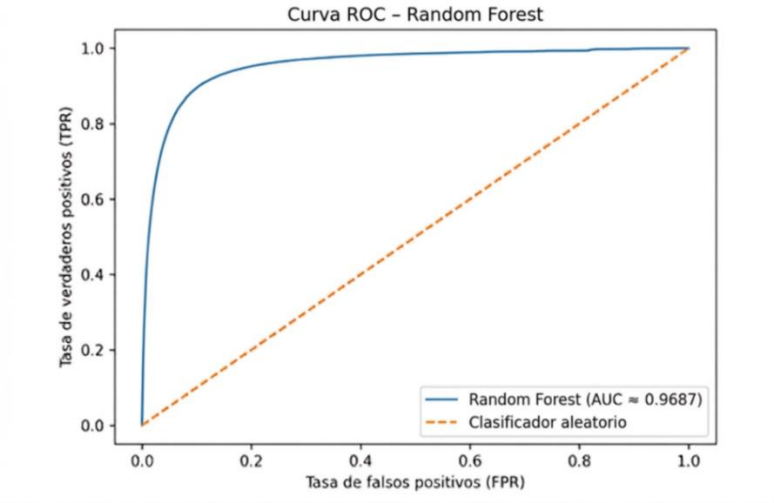


Figura 16. Curva ROC (BA)

Como se aprecia en la figura 16, el valor de ROC-AUC en test fue 0.9687. Este es el valor más alto entre los modelos implementados hasta el momento, superando al árbol de decisión (0.9494) y a la regresión logística (0.9508), lo que indica que el bosque aleatorio posee la mejor capacidad para separar probabilísticamente desertores y no desertores. Desde una perspectiva estadística, esto sugiere que el modelo captura interacciones complejas que no son completamente representadas por la regresión logística o por un árbol individual.

5.3.4. Análisis de importancia de variables

A diferencia del árbol de decisión simple donde una sola variable concentraba más del 80% de la importancia, el bosque aleatorio distribuye la importancia predictiva de manera más equilibrada entre múltiples variables, lo cual es apreciable en la tabla 15.

Tabla 15. Top 10 variables más influyentes en el modelo (BA)

Variable	Importancia (Gini)	Importancia relativa
Curricular units 2nd sem (approved)	0.2676	26.76%
Curricular units 1st sem (approved)	0.1780	17.80%
Curricular units 2nd sem (grade)	0.1385	13.85%
Curricular units 1st sem (grade)	0.0723	7.23%
Tuition fees up to date	0.0613	6.13%
Curricular units 2nd sem (evaluations)	0.0291	2.91%
Age at enrollment	0.0277	2.77%
Scholarship holder	0.0272	2.72%
Curricular units 1st sem (evaluations)	0.0226	2.26%
Course	0.0206	2.06%

El análisis revela una estructura multifactorial en la que el desempeño académico acumulado constituye el núcleo predictivo del modelo. Las cuatro variables de mayor importancia corresponden a unidades curriculares aprobadas y calificaciones en ambos semestres, concentrando en conjunto el 65.64% de la importancia total. Este patrón es consistente con la literatura sobre deserción estudiantil, que identifica el rendimiento académico temprano como el predictor más robusto del abandono.

La variable Curricular units 2nd sem (approved) lidera la importancia con un 26.76%, aunque su concentración es significativamente menor que en el árbol de decisión simple (80.53%), lo que refleja la capacidad del bosque aleatorio para distribuir el poder predictivo entre múltiples señales. En segundo lugar, Curricular units 1st sem (approved) aporta un 17.80%, reforzando que el número de materias aprobadas en ambos semestres constituye el núcleo predictivo del modelo.

Curricular units 2nd sem (grade) y Curricular units 1st sem (grade), con 13.85% y 7.23% respectivamente, confirman que el desempeño desde el primer semestre es un predictor

consistente del riesgo de deserción, incorporando además la dimensión del rendimiento cualitativo. La variable Tuition fees up to date (6.13%) incorpora la dimensión económica al modelo, indicando que el cumplimiento en el pago de matrícula es un factor diferenciador entre estudiantes que permanecen y los que abandonan.

Las variables restantes como evaluaciones, tipo de curso, beca, edad al matricularse, contribuyen individualmente con menos del 3%, pero en conjunto representan el 13.55% de la importancia total, evidenciando que el fenómeno de deserción responde a una estructura multicausal donde factores académicos, económicos, administrativos y sociodemográficos interactúan simultáneamente. Este hallazgo es coherente con el enfoque multivariado del presente estudio y respalda la pertinencia del bosque aleatorio como herramienta de modelado en contextos educativos complejos.

5.4. XGBOOST

Con el fin de evaluar un enfoque de aprendizaje supervisado con mayor capacidad de modelamiento no lineal, se implementó XGBoost (Extreme Gradient Boosting) para clasificación binaria. A diferencia del *bagging* (Random Forest), donde los árboles se entrenan de forma independiente y su principal efecto es la reducción de varianza, XGBoost aplica *boosting* secuencial: cada árbol corrige los errores del anterior mediante el ajuste del gradiente de la función de pérdida. Esta potencia predictiva puede incrementar el riesgo de modelos excesivamente ajustados si no se controlan adecuadamente los hiperparámetros, razón por la cual se incorporó una estrategia de selección robusta mediante validación cruzada.

5.4.1. Preparación de los datos

Se utilizó la base de datos previamente preparada y se aplicó one-hot encoding mediante pd.get_dummies() sobre las variables categóricas, conforme a lo descrito en el Capítulo I. Posteriormente, el conjunto se dividió en entrenamiento (80%) y prueba (20%) mediante partición estratificada, preservando la proporción de clases.

Dado el desbalance entre clases, se implementó un ajuste explícito a través del parámetro: scale_pos_weight = neg/pos. Donde neg corresponde al número de observaciones de la clase 0 y pos al número de observaciones de la clase 1 dentro del conjunto de entrenamiento. Este ajuste modifica la función objetivo para penalizar proporcionalmente los errores en la clase minoritaria, favoreciendo una sensibilidad más equilibrada.

5.4.2. Optimización de hiperparámetros

Se realizó una búsqueda en malla mediante GridSearchCV con validación cruzada estratificada repetida (5 folds × 3 repeticiones), para un total de 60 combinaciones y 900 ajustes, empleando como criterio principal el F1-score para equilibrar precisión y recall en la clase de desertores. Los parámetros de regularización y muestreo (subsample = 0.8,

colsample_bytree = 0.8, gamma = 0.0, reg_alpha = 0.0, reg_lambda = 1.0) se fijaron para restringir el espacio de búsqueda y mantener consistencia experimental. La tabla siguiente presenta los hiperparámetros explorados y el óptimo seleccionado

Tabla 16. Búsqueda y selección de hiperparámetros XGBoost

Hiperparámetro	Valores evaluados	Valor óptimo
n_estimators	200, 350, 500	500
learning_rate	0.03, 0.06	0.03
max_depth	3, 4, 5, 6, 7	4
min_child_weight	3, 7	3
subsample	Regularización	0.8
colsample_bytree	Regularización	0.8
reg_lambda	Regularización	1.0

El mejor F1-score promedio en validación cruzada fue F1CV = 0.8705. Este valor supera al árbol de decisión simple (0.8415) y es comparable al obtenido con Random Forest (0.8661), lo cual es consistente con la capacidad de los métodos de *boosting* para mejorar el ajuste mediante aprendizaje secuencial. La selección de una tasa de aprendizaje baja (learning_rate = 0.03) junto con una profundidad máxima de 4 refleja un balance adecuado entre capacidad expresiva y control del sobreajuste.

5.4.3. Evaluación del desempeño y generalización

Una vez seleccionado el modelo óptimo, se evaluó su desempeño durante la fase de optimización mediante validación cruzada estratificada repetida, y posteriormente sobre los conjuntos de entrenamiento y prueba. Las métricas obtenidas se presentan en las tablas siguientes

Tabla 17. Métricas promedio en validación cruzada XGBoost

Métrica	Media CV (5 folds x 3 repeticiones)	Desv. Estándar
Accuracy	0.9020	± 0.0132
Precision	0.9014	± 0.0208
Recall	0.8420	± 0.0216
F1-score	0.8705	± 0.0177
ROC-AUC	0.9498	± 0.0101

Tabla 18. Métricas de desempeño del modelo XGBoost

Métrica	Entrenamiento	Prueba	Diferencia absoluta
Accuracy	0.9649	0.9201	0.0448

Precision	0.9675	0.8897	0.0778
Recall	0.9420	0.9085	0.0335
F1-score	0.9545	0.8990	0.0555
ROC-AUC	0.9948	0.9728	0.0220

Los resultados de la validación cruzada (Tabla 17) evidencian un desempeño estable durante la fase de optimización, con desviaciones estándar bajas en todas las métricas (≤ 0.0216), confirmando que la selección de hiperparámetros no dependió de una partición particular de los datos. En el conjunto de prueba, el modelo alcanza un F1-score de 0.8990 y un ROC-AUC de 0.9728, evidenciando un desempeño sobresaliente y una alta capacidad discriminativa. Las brechas Train-Test son mayores que en Random Forest, especialmente en precisión (0.0778) y F1 (0.0555), lo cual es coherente con la naturaleza de XGBoost: al optimizar secuencialmente la función de pérdida, tiende a ajustarse con mayor detalle al conjunto de entrenamiento, incrementando la brecha esperable frente a datos no vistos. No obstante, el desempeño en prueba se mantiene elevado, lo que sugiere que el modelo conserva una adecuada capacidad de generalización, gracias a la tasa de aprendizaje baja (learning_rate = 0.03) que ralentiza el ajuste y reduce el riesgo de memorización, el muestreo de filas y columnas (subsample = 0.8, colsample_bytree = 0.8) que introduce aleatoriedad y reduce la correlación entre árboles, y el ajuste de clase (scale_pos_weight) que equilibra la contribución de ambas clases a la función de pérdida.

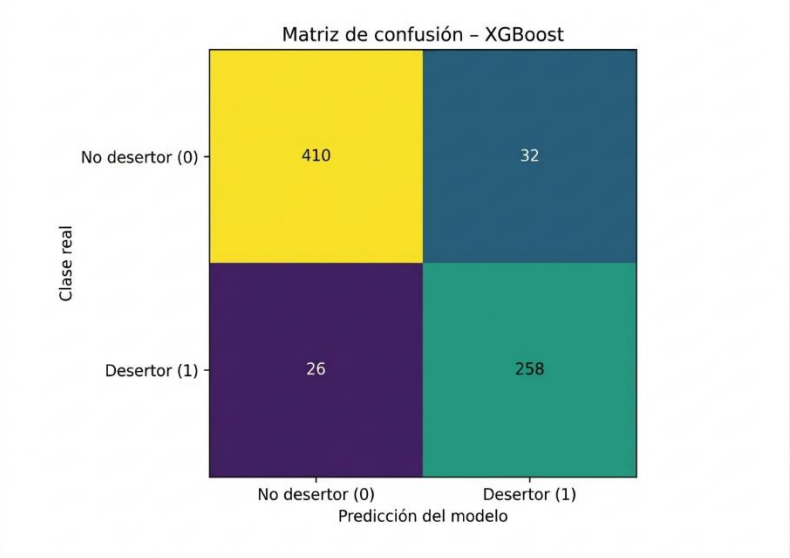


Figura 17. Matriz de confusión XGBoost

El modelo clasifica correctamente 410 estudiantes no desertores (verdaderos negativos) y 258 desertores (verdaderos positivos), registrando 32 falsos positivos y 26 falsos negativos.

Lo anterior representa una sensibilidad (Recall clase 1) del 90.85% y una especificidad (clase 0) del 92.76%.

En comparación con el bosque aleatorio, XGBoost reduce los falsos negativos de 30 a 26, lo cual es particularmente valioso en el contexto institucional: los falsos negativos representan estudiantes con riesgo real que no serían priorizados por estrategias preventivas. El recall de 0.9085 evidencia un desempeño alto en la identificación efectiva de desertores, sin deteriorar significativamente la precisión (0.8897), lo que implica que la priorización de casos para intervención no se incrementa artificialmente por exceso de alertas.

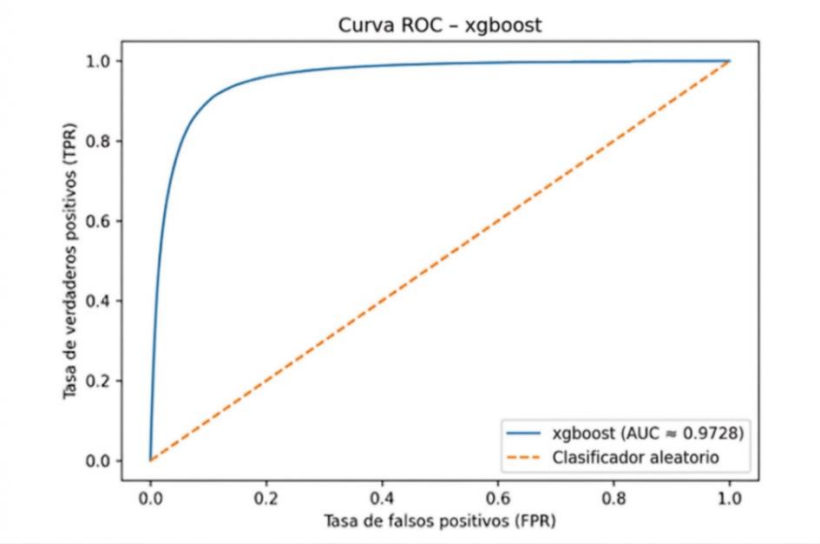


Figura 18. Curva ROC XGBoost

El valor de ROC-AUC en prueba fue 0.9728, el más alto registrado entre todos los modelos evaluados hasta el momento, lo que indica una separación probabilística muy robusta entre desertores y no desertores.

5.4.4. Análisis de importancia de variables

La importancia de variables en XGBoost se midió mediante el criterio gain, que cuantifica la reducción promedio de la función de pérdida aportada por cada variable en todas las divisiones donde interviene. A diferencia de la importancia por frecuencia (que solo cuenta cuántas veces se usa una variable), el gain refleja la contribución efectiva al poder predictivo del modelo. La tabla 19 presenta las diez variables con mayor importancia.

Tabla 19. Top 10 variables más influyentes en XGBoost

Variable	Importancia (gain)	Importancia relativa
Curricular units 2nd sem (approved)	0.2386	23.86%
Tuition fees up to date	0.0904	9.04%
Curricular units 1st sem (approved)	0.0719	7.19%
Curricular units 1st sem (enrolled)	0.0395	3.95%
Debtor	0.0357	3.57%
Scholarship holder	0.0320	3.20%
Curricular units 2nd sem (enrolled)	0.0315	3.15%
Curricular units 2nd sem (evaluations)	0.0249	2.49%
Curricular units 2nd sem (grade)	0.0245	2.45%
Course	0.0242	2.42%

La variable Curricular units 2nd sem (approved) lidera con un 23.86% de importancia, confirmando un patrón consistente con los modelos anteriores: el número de materias aprobadas en el segundo semestre es el predictor más robusto de la deserción estudiantil. Su concentración en XGBoost (23.86%) es intermedia entre la del árbol de decisión simple (80.53%) y la del bosque aleatorio (26.76%), lo que refleja la capacidad del *boosting* para diversificar el uso de variables sin diluir la señal principal.

Tuition fees up to date ocupa el segundo lugar con un 9.04%, evidenciando que el cumplimiento en el pago de matrícula constituye una señal financiera de alto valor predictivo. A diferencia del bosque aleatorio, donde las variables de unidades curriculares aprobadas ocupaban los primeros lugares, XGBoost asigna mayor importancia a esta variable económica, lo que sugiere que el algoritmo identifica en el estado de pago un punto de quiebre temprano y eficiente para discriminar entre desertores y no desertores.

Curricular units 1st sem (approved) contribuye con un 7.19%, reforzando la relevancia del desempeño académico desde el primer semestre como predictor anticipado del riesgo de abandono. Le siguen Curricular units 1st sem (enrolled) (3.95%) y Debtor (3.57%), que incorporan dimensiones de carga académica y económica respectivamente, señalando que tanto la cantidad de materias inscritas, como la situación de deuda están asociadas al riesgo de deserción.

Las cinco variables restantes del top 10 aportan individualmente entre el 2.42% y el 3.20%, pero en conjunto representan el 16.71% de la importancia total. Este patrón confirma la naturaleza multifactorial del fenómeno: aunque el desempeño académico temprano domina la predicción, factores económicos, administrativos y de carga curricular interactúan de manera complementaria para explicar el abandono estudiantil.

En comparación con Random Forest, XGBoost muestra una mayor concentración en un subconjunto reducido de variables académicas y financieras, lo que puede interpretarse como evidencia de que el modelo identifica variables de alto impacto en la reducción de pérdida durante el proceso de *boosting* secuencial. No obstante, se reconoce que la interpretación de importancias en modelos de *boosting* debe considerarse como un indicador relativo de utilidad predictiva y no como una medida de causalidad.

5.5. RED NEURONAL TOTALMENTE CONECTADA

Con el propósito de incorporar un enfoque de aprendizaje supervisado con alta capacidad para capturar patrones no lineales y combinaciones complejas entre variables, se implementó una red neuronal artificial totalmente conectada (*fully conected*). A diferencia de modelos basados en árboles, una red neuronal aprende una representación jerárquica del fenómeno a través de transformaciones sucesivas del espacio de características, permitiendo modelar relaciones no lineales sin requerir especificación explícita de interacciones.

En el contexto de la predicción de deserción estudiantil, las redes neuronales pueden ser particularmente útiles cuando la decisión final depende de combinaciones de variables académicas, económicas y administrativas cuya relación no es estrictamente aditiva o jerárquica. No obstante, por su elevada flexibilidad, las redes neuronales requieren estrategias explícitas de control de complejidad y estabilidad del entrenamiento (regularización, *early stopping*, escalado y control de aleatoriedad), las cuales fueron incorporadas en el presente estudio.

5.5.1. Preparación de los datos

Se utilizó la base utilizada en los modelos anteriores, separando la variable objetivo (Target) de la matriz de predictores. Se aplicó one-hot encoding mediante pd.get_dummies() sobre las variables categóricas, conforme a lo descrito en el Capítulo I, garantizando representaciones binarias aptas para el entrenamiento.

A diferencia de modelos basados en árboles (que no son sensibles a escalas), las redes neuronales requieren que las variables estén en escalas comparables para estabilizar gradientes y acelerar convergencia. Por ello se aplicó estandarización (StandardScaler), ajustada únicamente con el conjunto de entrenamiento y posteriormente aplicada al conjunto de prueba, reduciendo el riesgo de fuga de información.

La partición entrenamiento/prueba se realizó con proporción 80/20 y estratificación para preservar la proporción de clases. Asimismo, se fijaron semillas (NumPy y TensorFlow) con el fin de garantizar reproducibilidad experimental.

5.5.2. Arquitectura y estrategia de regularización

La arquitectura definida correspondió a un modelo secuencial con dos capas ocultas totalmente conectadas:

- Capa 1: 64 neuronas, activación ReLU, regularización L2 (1e-4)
  Dropout: 0.3
- Capa 2: 32 neuronas, activación ReLU, regularización L2 (1e-4)
  Dropout: 0.2
- Capa de salida: 1 neurona con activación sigmoide (clasificación binaria)

Los hiperparámetros de entrenamiento se configuraron como se indica en la tabla siguiente.

Tabla 20. Hiperparámetros de entrenamiento (RN)

Parámetro	**Configuración**
Optimizador	Adam
Learning rate	0.001
Función de pérdida	Binary crossentropy
Épocas máximas	100
Batch size	256
Validation split	0.20
Early stopping (patience)	10 épocas (val_loss)

El control del sobreajuste se implementó mediante tres mecanismos complementarios:

- Regularización L2 (1e-4): penaliza pesos grandes y favorece soluciones más suaves, reduciendo la capacidad del modelo para memorizar patrones espurios del entrenamiento.
- Dropout (0.3 en capa 1, 0.2 en capa 2): desactiva neuronas aleatoriamente durante el entrenamiento, reduciendo coadaptaciones y actuando como un mecanismo implícito de promediado de modelos.
- Early Stopping (patience=10, monitor=val_loss): detiene el entrenamiento cuando la pérdida en validación deja de mejorar por 10 épocas consecutivas, restaurando los mejores pesos. Este mecanismo reemplaza funcionalmente al esquema de validación cruzada utilizado en los modelos anteriores.

5.5.3. Evaluación del desempeño y generalización

El desempeño del modelo se evaluó en el conjunto de entrenamiento y en el conjunto de prueba. Las métricas obtenidas, junto con la brecha absoluta y el diagnóstico de sobreajuste, se presentan en la tabla siguiente.

Tabla 21. Métricas de desempeño del modelo (RN)

Métrica	Entrenamiento	Prueba	Diferencia absoluta
Accuracy	0.9163	0.9435	0.0272
Precision	0.9488	0.9517	0.0029
Recall	0.8311	0.9014	0.0703
F1-score	0.8861	0.9259	0.0398
ROC-AUC	0.9622	0.9686	0.0064

Los resultados evidencian un desempeño sobresaliente en el conjunto de prueba, con un F1-score de 0.9259 y un ROC-AUC de 0.9686. Estos valores superan a la regresión logística, al árbol de decisión, al bosque aleatorio y a XGBoost en F1, posicionando a la red neuronal como el mejor modelo en esta métrica, aunque XGBoost mantiene superioridad en Recall (0.9085 vs 0.9014).

Un aspecto metodológicamente relevante es que varias métricas son superiores en prueba que en entrenamiento (Accuracy: 0.9435 vs 0.9163; Recall: 0.9014 vs 0.8311; F1: 0.9259 vs 0.8861). Este comportamiento, aunque inusual, es explicable por la acción conjunta de los mecanismos de regularización: el *early stopping* y el *dropout* limitan deliberadamente la capacidad del modelo para ajustarse al conjunto de entrenamiento, produciendo un modelo más conservador que generaliza mejor de lo que memoriza. La baja diferencia en ROC-AUC (0.0064) confirma estabilidad en la capacidad discriminativa entre ambas particiones.

Desde la perspectiva institucional, es importante destacar que el Recall en prueba (0.9014) es el indicador más crítico para la toma de decisiones, ya que mide la proporción de desertores reales que el modelo logra identificar. Un falso negativo representa un estudiante en riesgo que no sería priorizado por estrategias de intervención, lo que implica una oportunidad perdida de retención. En este sentido, aunque la red neuronal obtiene la mayor Precision de todos los modelos (0.9517), su Recall es superado por XGBoost (0.9085), lo que debe considerarse al momento de seleccionar el modelo para implementación institucional.

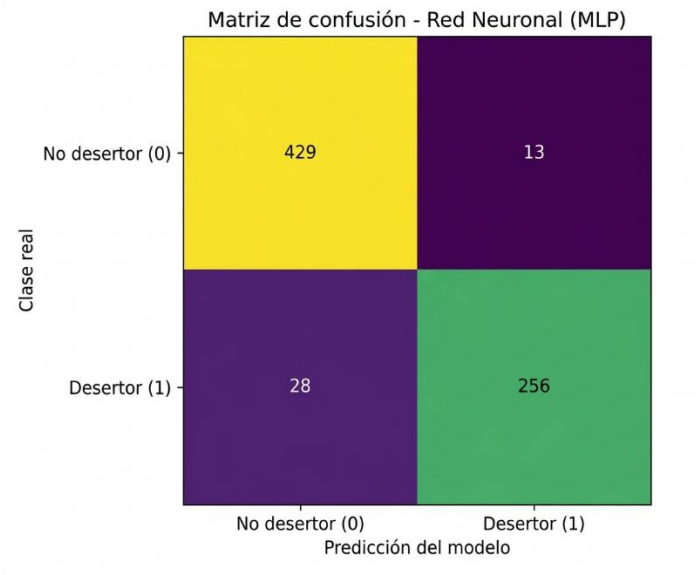


Figura 19. Matriz de confusión (RN)

El modelo clasifica correctamente 429 estudiantes no desertores (verdaderos negativos) y 256 desertores (verdaderos positivos), registrando 13 falsos positivos y 28 falsos negativos. Lo anterior representa una sensibilidad (Recall clase 1) del 90.14% y una especificidad (clase 0) del 97.06%. El resultado más destacable es la reducción de falsos positivos a 13, el valor más bajo entre todos los modelos evaluados. Esto implica que la red neuronal genera el menor número de alertas incorrectas, lo que desde una perspectiva operativa se traduce en un uso más eficiente de los recursos institucionales destinados a intervención. No obstante, los 28 falsos negativos son ligeramente superiores a los de XGBoost (26), lo que confirma que existe una tensión inherente entre Precision y Recall: maximizar una tiende a reducir la otra.

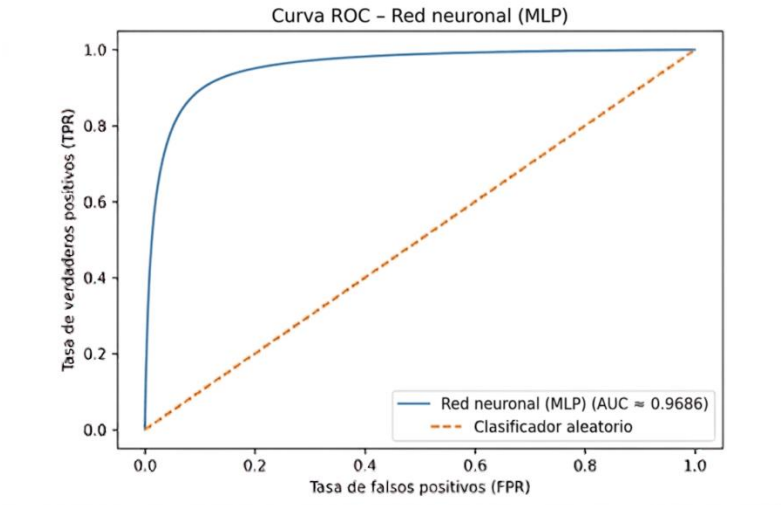


Figura 20. Curva ROC (RN)

El valor de ROC-AUC en prueba fue 0.9686, evidenciando una alta capacidad discriminativa. Este valor es ligeramente inferior al de XGBoost (0.9728) pero superior al de Random Forest (0.9687), aunque la diferencia es marginal. En términos absolutos, los tres modelos presentan una capacidad discriminativa muy similar y estadísticamente comparable.

5.5.4. Análisis de importancia de variables

Dado que una red neuronal no ofrece interpretabilidad directa como los coeficientes de la regresión logística o las reglas de los modelos basados en árboles, se implementó un análisis post-hoc basado en Permutation Importance. Este método mide la caída en ROC-AUC al permutar aleatoriamente cada variable de forma individual sobre el conjunto de prueba: una caída grande indica que la variable es altamente relevante para el modelo, mientras que una caída cercana a cero sugiere escasa contribución predictiva. La tabla siguiente presenta las diez variables con mayor importancia.

Tabla 22. Top 10 variables más influyentes (RN)

Variable	Importancia (caída AUC)	Importancia relativa
Curricular units 2nd sem (approved)	0.0997	36.73%
Curricular units 1st sem (approved)	0.0444	16.34%
Tuition fees up to date	0.0238	8.75%
Curricular units 2nd sem (grade)	0.0198	7.28%
Curricular units 2nd sem (credited)	0.0132	4.86%
Course	0.0115	4.25%
Curricular units 2nd sem (enrolled)	0.0070	2.59%
Curricular units 1st sem (credited)	0.0069	2.56%
Debtor	0.0068	2.50%
Scholarship holder	0.0067	2.46%

La variable Curricular units 2nd sem (approved) lidera con una importancia relativa del 36.73%, confirmando el patrón consistente observado en todos los modelos implementados: el número de materias aprobadas en el segundo semestre es el predictor más robusto del riesgo de deserción, independientemente del algoritmo empleado. Esta consistencia inter-modelo fortalece la validez del hallazgo y lo convierte en un insumo de alto valor para la toma de decisiones institucionales.

Curricular units 1st sem (approved) ocupa el segundo lugar con un 16.34%, evidenciando que el desempeño académico desde el primer semestre también constituye una señal temprana y confiable de riesgo. En conjunto, las dos variables de unidades aprobadas concentran el 53.07% de la importancia total, subrayando el papel central del rendimiento académico acumulado en la predicción de la deserción.

Tuition fees up to date (8.75%) y Debtor (2.50%) incorporan la dimensión económica al modelo, confirmando que el cumplimiento en el pago de matrícula y la condición de deuda son factores diferenciadores relevantes. La presencia de Course (4.25%) entre las variables de mayor importancia introduce una dimensión programática, sugiriendo que el programa académico cursado también modula el riesgo de abandono, posiblemente por diferencias en exigencia académica, empleabilidad percibida o perfil socioeconómico del estudiantado.

En contraste con XGBoost, donde Debtor aparecía con mayor peso relativo (3.57%), la red neuronal asigna mayor importancia a las calificaciones del segundo semestre (7.28%) y las unidades curriculares acreditadas (4.86% y 2.56%), lo que refleja que cada modelo captura matices distintos de la misma realidad multifactorial. Esta diversidad de perspectivas es, en sí misma, un argumento a favor del enfoque comparativo adoptado en el presente estudio.

En conjunto, el patrón de importancias de la red neuronal es coherente con los hallazgos de los modelos anteriores: el fenómeno de deserción estudiantil responde a una estructura multifactorial en la que el desempeño académico temprano, la regularidad financiera y factores administrativos y programáticos interactúan de manera complementaria. Esta consistencia refuerza la robustez de los resultados y respalda la pertinencia de las variables seleccionadas para la predicción del abandono estudiantil.

5.6. SELECCIÓN DEL MODELO FINAL

Con el objetivo de identificar el modelo con mejor desempeño para la predicción de deserción estudiantil, se realizó una comparación sistemática entre los cinco modelos evaluados. Todas las métricas reportadas corresponden al conjunto de prueba (Test), garantizando que la comparación refleje la capacidad de generalización de cada modelo ante datos no vistos durante el entrenamiento.

Con el fin de garantizar la objetividad y la trazabilidad del proceso de selección, se estableció de manera explícita la siguiente regla de decisión jerárquica, consistente con el criterio de priorización del Recall ya enunciado en la introducción del Capítulo II: (1) se selecciona como modelo final aquel que presente el mayor Recall en el conjunto de prueba, dado que esta métrica refleja directamente la capacidad de detección de estudiantes en riesgo real de deserción; (2) en caso de empate o diferencias no sustantivas en Recall, se utiliza el ROC-AUC como criterio de desempate, por representar la capacidad discriminativa global del modelo independientemente del umbral de clasificación. Esta regla fue definida con anterioridad a la comparación de resultados, evitando así que la elección del modelo final responda a un análisis realizado después de conocer los resultados con el fin de justificar una preferencia particular.

Tabla 23. Comparación del desempeño de los modelos en el conjunto de prueba

Modelo	Accuracy	Precision	Recall	F1-score	ROC-AUC
Regresión logística	0.9036	0.8905	0.8592	0.8746	0.9508
Árbol de decisión	0.9022	0.8790	0.8697	0.8743	0.9494
Random Forest	0.9132	0.8850	0.8944	0.8897	0.9687
XGBoost *	**0.9201**	**0.8897**	**0.9085**	**0.8990**	**0.9728**
Red neuronal (MLP)	0.9435	0.9517	0.9014	0.9259	0.9686

* Modelo seleccionado como modelo final. Fila destacada por mayor Recall y ROC-AUC.

La tabla 23 presenta las métricas de desempeño obtenidas en el conjunto de prueba para los cinco modelos implementados. Los métodos de ensamble (Random Forest y XGBoost) y la red neuronal superan a la regresión logística y al árbol de decisión en la mayoría de las métricas, evidenciando que su capacidad para modelar relaciones no lineales entre las variables predictoras se traduce en una mayor discriminación entre estudiantes desertores y no desertores.

5.6.1. Análisis por métrica de desempeño

Dado que el objetivo del sistema predictivo es apoyar estrategias institucionales de retención estudiantil, la métrica de mayor relevancia en este contexto es la sensibilidad (recall), que mide la proporción de desertores reales que el modelo logra identificar correctamente. Un falso negativo representa un estudiante en riesgo que no sería priorizado por las estrategias de intervención, lo que implica una oportunidad perdida de retención con alto costo institucional y social. Por tanto, maximizar el Recall es prioritario frente a maximizar la precisión.

Desde esta perspectiva, XGBoost obtiene el Recall más alto entre todos los modelos evaluados (0.9085), identificando correctamente al 90.85% de los estudiantes desertores en el conjunto de prueba. La tabla 24 presenta la distribución de verdaderos positivos, falsos positivos, falsos negativos y verdaderos negativos para cada modelo, permitiendo comparar directamente el impacto institucional de los errores de clasificación.

Tabla 24. Comparación de matrices de confusión en el conjunto de prueba

Modelo	VP	FP	FN	VN
Regresión logística	244	30	40	412
Árbol de decisión	247	34	37	408
Random Forest	254	33	30	409
XGBoost *	**258**	**32**	**26**	**410**
Red neuronal (MLP)	256	13	28	429

VP: verdaderos positivos (desertores correctamente identificados). FP: falsos positivos (no desertores clasificados como desertores). FN: falsos negativos (desertores no identificados). VN: verdaderos negativos.

XGBoost registra el menor número de falsos negativos (26), seguido de la red neuronal (28) y Random Forest (30). La regresión logística presenta el mayor número de casos no detectados (40), mientras que el árbol de decisión (37) se ubica en una posición intermedia. Este análisis confirma que los modelos de mayor complejidad ofrecen ventajas concretas en la identificación de estudiantes en riesgo, con implicaciones directas para la efectividad de los programas de permanencia estudiantil.

5.6.2. Análisis comparativo global

En cuanto a la capacidad discriminativa global, XGBoost presenta el valor más alto de ROC-AUC (0.9728), lo que indica su mayor habilidad para separar probabilísticamente desertores y no desertores a lo largo de diferentes umbrales de decisión. Random Forest (0.9687) y la red neuronal (0.9686) le siguen con valores numéricamente muy próximos, evidenciando que los tres modelos poseen una capacidad discriminativa similar y sobresaliente.

La red neuronal (MLP) obtiene el mayor F1-score (0.9259) y la mayor exactitud (0.9435), así como la mayor precision (0.9517), lo que se traduce en el menor número de falsos positivos (13) entre todos los modelos. Esto implica que cuando la red neuronal alerta sobre un posible desertor, lo hace con alta certeza, reduciendo el riesgo de intervenciones innecesarias. Sin embargo, su Recall (0.9014) es inferior al de XGBoost, lo que significa que deja escapar más estudiantes en riesgo real.

Esta tensión entre Precision y Recall es inherente a cualquier clasificador binario y refleja un dilema institucional concreto: ¿es más costoso intervenir sobre estudiantes que no necesitan apoyo (falsos positivos), o no intervenir sobre estudiantes que sí lo necesitan (falsos negativos)? En el contexto de la deserción estudiantil, donde el costo del abandono en términos humanos, sociales y financieros supera al costo de una intervención preventiva innecesaria, minimizar los falsos negativos es la prioridad estratégica.

Es importante señalar que las comparaciones presentadas se basan en las estimaciones puntuales de las métricas de desempeño sobre el conjunto de prueba, sin la aplicación de pruebas de significancia estadística pareadas (p. ej., test de McNemar o de DeLong) que permitan determinar si las diferencias observadas entre modelos son estadísticamente significativas o si podrían atribuirse a la variabilidad muestral del conjunto de prueba. Esta decisión responde, en parte, a una limitación metodológica del diseño experimental: dado que el esquema de validación adoptado para la regresión logística utilizó una semilla de partición distinta (*random_state* = 12) a la empleada en los demás modelos (*random_state*

= 42) por razones de estabilidad estadística discutidas previamente, una comparación pareada rigurosa habría requerido homogeneizar las particiones de prueba entre todos los modelos, lo cual excedía el alcance definido para el presente estudio. Se reconoce esta omisión como una oportunidad de mejora metodológica, la cual se retoma en la sección de trabajos futuros.

5.6.3. Selección del modelo final

Considerando el conjunto de evidencia presentado, XGBoost fue seleccionado como modelo final para su implementación en el sistema de monitoreo de riesgo de deserción estudiantil, con base en los siguientes criterios:

- Mayor Recall en prueba (0.9085): identifica al 90.85% de los estudiantes en condición de deserción, minimizando los falsos negativos. Esta métrica es priorizada institucionalmente dado que el costo de no detectar un estudiante en riesgo (falso negativo) es significativamente mayor que el de generar una alerta innecesaria (falso positivo).
- Mayor ROC-AUC en prueba (0.9728): posee la mejor capacidad discriminativa global, manteniendo un alto desempeño a lo largo de diferentes umbrales de decisión.
- Desempeño consistente en todas las métricas: Accuracy (0.9201), Precision (0.8897) y F1-score (0.8990) se mantienen en niveles altos, evidenciando un modelo balanceado entre detección de desertores y precisión predictiva.
- Brechas Train-Test controladas: las diferencias entre desempeño en entrenamiento y prueba se mantienen dentro de rangos aceptables, gracias a los mecanismos de regularización incorporados al modelo (tasa de aprendizaje baja, muestreo de filas y columnas, ajuste por clase), lo que respalda la capacidad de generalización del modelo final.
- Interpretabilidad mediante importancia de variables: XGBoost permite cuantificar de forma nativa la contribución relativa de cada variable mediante el criterio gain, que mide la mejora promedio en la función de pérdida cuando una variable es utilizada en las particiones de los árboles. Esta característica facilita identificar los factores más determinantes de la deserción y traducir los resultados en información accionable para los procesos institucionales de seguimiento y retención.

Si bien la red neuronal supera a XGBoost en F1-score (0.9259 vs 0.8990), Accuracy (0.9435 vs 0.9201) y precision (0.9517 vs 0.8897), y constituye una alternativa de alto valor predictivo, la prioridad institucional de minimizar falsos negativos, es decir estudiantes en riesgo no detectados, inclina la balanza hacia XGBoost como la opción más adecuada para el objetivo específico del presente estudio. No obstante, se recomienda que futuras

implementaciones evalúen ajustes en el umbral de clasificación de la red neuronal, lo que podría mejorar su Recall sin sacrificar significativamente su elevada precisión.

6. CAPITULO III. IMPLEMENTAR UN PROTOTIPO FUNCIONAL QUE PERMITA VISUALIZAR LAS PREDICCIONES REALIZADAS POR EL MEJOR MODELO

El presente capítulo describe el desarrollo e implementación de un prototipo funcional orientado a la visualización interactiva de las predicciones generadas por el modelo de clasificación XGBoost. Considerándolo una herramienta de apoyo frente a la toma de decisiones institucionales, al permitir identificar de forma temprana a los estudiantes con mayor probabilidad de deserción en una institución de educación superior.

El sistema fue desarrollado utilizando el marco de trabajo Streamlit, un framework de código abierto en Python diseñado para la construcción rápida de aplicaciones web interactivas orientadas a la ciencia de datos y el aprendizaje automático [45]. La aplicación fue desplegada en Streamlit Community Cloud, lo que garantiza su disponibilidad permanente y accesibilidad desde cualquier dispositivo con conexión a Internet, sin depender de entornos de ejecución locales.

La arquitectura del prototipo separa explícitamente la fase de entrenamiento del modelo de la fase de inferencia y visualización. El modelo entrenado es serializado y almacenado como un artefacto independiente en formato .pkl, junto con los metadatos necesarios para realizar predicciones sobre nuevos datos sin necesidad de reentrenamiento. Esta decisión de diseño garantiza la reproducibilidad de las predicciones y reduce significativamente los tiempos de respuesta del sistema.

6.1. ARQUITECTURA DEL SISTEMA

El sistema implementado sigue una arquitectura de dos fases claramente diferenciadas: una fase de entrenamiento y serialización del modelo, y una fase de despliegue e inferencia sobre datos nuevos. Esta separación es fundamental para garantizar que el modelo utilizado en producción sea exactamente el mismo que fue optimizado mediante el proceso de búsqueda de hiperparámetros descrito en el Capítulo II.

Para facilitar la comprensión de la interacción entre los distintos componentes tecnológicos y el movimiento de los datos, a continuación, se presenta la Figura 21. En este esquema se detalla el flujo integral del sistema, desde la ingesta del dataset histórico y el entrenamiento del modelo XGBoost en el backend, hasta la generación de los cinco artefactos serializados que alimentan la interfaz de usuario en el frontend. El diagrama ilustra cómo el repositorio de GitHub actúa como puente de auto-despliegue hacia Streamlit Community Cloud, permitiendo que el procesamiento de inferencia y la visualización de métricas ocurran de manera sincronizada y eficiente.

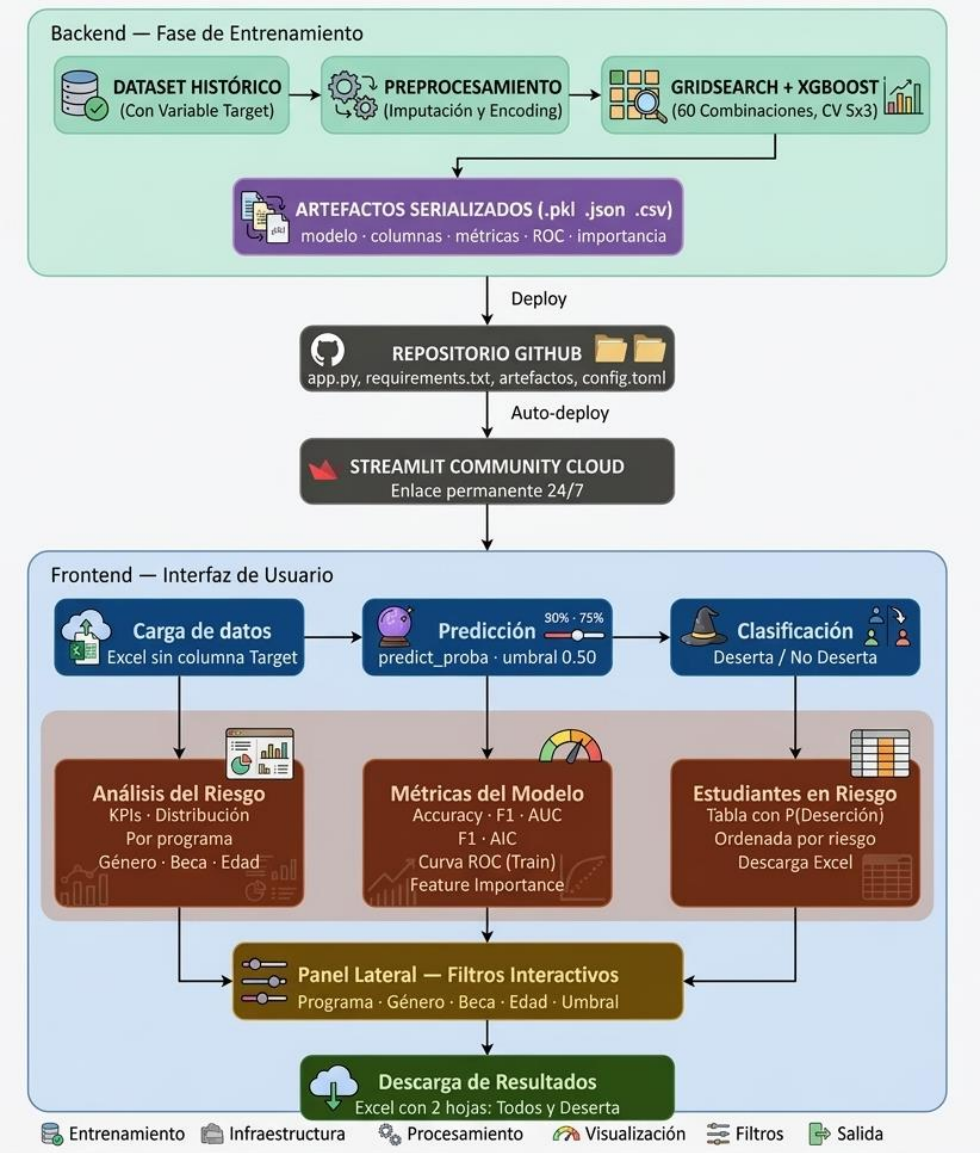


Figura 21. Arquitectura general del prototipo de visualización

6.1.1. Fase de entrenamiento y serialización

Una vez identificado el mejor modelo a través del proceso de validación cruzada repetida (RepeatedStratifiedKFold con k=5 y 3 repeticiones) y la búsqueda en rejilla (GridSearchCV) sobre 60 combinaciones de hiperparámetros, el modelo resultante se serializó mediante la librería joblib [46]. Este proceso genera cinco artefactos que son necesarios para el funcionamiento del prototipo:

- modelo_desercion.pkl: archivo binario que contiene el clasificador XGBoost entrenado con los hiperparámetros óptimos.

- columnas_modelo.json: lista con los nombres exactos de las columnas utilizadas durante el entrenamiento, necesaria para alinear correctamente los datos de entrada en la fase de inferencia.
- metricas_train.json: métricas de desempeño del modelo sobre el conjunto de entrenamiento (Accuracy, Precision, Recall, F1-Score y ROC AUC).
- roc_train.json: datos de la curva ROC del conjunto de entrenamiento para su visualización en el dashboard.
- feature_importance.csv: importancia de cada variable según el criterio de ganancia (gain) del modelo XGBoost.

6.1.2. Fase de inferencia y visualización

En la fase de producción, el prototipo carga los artefactos serializados al iniciar la sesión y los almacena en caché para evitar recargas innecesarias. El usuario final únicamente proporciona un archivo Excel con los datos de los estudiantes a evaluar, sin incluir la variable objetivo (Target), ya que el sistema genera las predicciones de forma automática. El flujo de procesamiento comprende los siguientes pasos: (1) carga y validación del archivo de entrada, (2) aplicación de codificación de variables categóricas mediante one-hot encoding, (3) alineación de columnas con el esquema del modelo entrenado, (4) generación de probabilidades de deserción mediante predict_proba, y (5) clasificación binaria de cada estudiante según el umbral definido por el usuario, el cual por defecto es 0.5.

6.2. DESCRIPCIÓN DEL PROTOTIPO

El prototipo implementado recibe el nombre de Sistema de Predicción de Deserción en Estudiantes de Educación Superior. La interfaz fue diseñada con un esquema visual de tema oscuro para mejorar la legibilidad en entornos de trabajo institucional, empleando una paleta de colores semafórica que diferencia visualmente las categorías de riesgo: verde para los estudiantes clasificados como "No Deserta" y rojo para los clasificados como "Deserta".

La aplicación se estructura en tres módulos principales, accesibles a través de un sistema de pestañas: (1) Análisis del Riesgo, (2) Métricas del Modelo, y (3) Estudiantes en Riesgo. Adicionalmente, dispone de un panel lateral de filtros interactivos y una sección introductoria con instrucciones de uso.

6.2.1. Pantalla principal e instrucciones

La pantalla principal del sistema presenta el título del proyecto, una descripción del modelo subyacente y una sección expandible de instrucciones de uso. Esta sección orienta al usuario sobre el proceso de carga de datos, describiendo el formato requerido del archivo excel de entrada y poniendo a disposición una plantilla descargable que incluye la descripción de cada variable y un ejemplo de registro válido. El botón de descarga de la plantilla se conecta directamente al archivo de referencia almacenado en el repositorio, garantizando que el usuario cuente siempre con el formato actualizado.

Figura 22. Pantalla principal del sistema de predicción de deserción

6.2.2. Panel lateral de filtros

El panel lateral izquierdo permite al usuario segmentar la población de estudiantes analizada mediante cuatro filtros interactivos: programa académico, género, condición de beca y rango de edad. Adicionalmente, incluye un control deslizante para ajustar el umbral de clasificación, cuyo valor por defecto es 0.50, en concordancia con el punto de corte estándar en clasificación binaria. Modificar el umbral recalcula en tiempo real la clasificación de todos los estudiantes, permitiendo explorar escenarios de mayor sensibilidad (umbral bajo) o mayor especificidad (umbral alto) según las necesidades institucionales.

6.3. MODULO 1: ANÁLISIS DE RIESGO

El primer módulo constituye el componente central del prototipo y concentra el conjunto de visualizaciones diseñadas para el análisis exploratorio de las predicciones. Se organiza en cuatro secciones: indicadores estratégicos, distribución de la clasificación, deserción por programa académico, riesgo por variables demográficas, y análisis por edad.

6.3.1. Indicadores

En la parte superior del módulo se presentan cuatro indicadores de resumen (KPIs): el número total de estudiantes analizados, el conteo y porcentaje de estudiantes clasificados como "Deserta", el conteo y porcentaje de estudiantes clasificados como "No Deserta", y la probabilidad promedio de deserción de la población filtrada. Estos indicadores se actualizan dinámicamente en respuesta a los filtros aplicados desde el panel lateral.

Debajo de los indicadores se presentan dos gráficas complementarias: un gráfico de barras que muestra el conteo absoluto de estudiantes por categoría de clasificación, y un gráfico de anillo que expresa las proporciones relativas de cada categoría. En el ejemplo ilustrado en la Figura 23, sobre una población de 1.000 estudiantes, el modelo clasificó a 330 (33%) como probables desertores y a 670 (67%) como estudiantes que continuarán sus estudios, con una probabilidad promedio de deserción del 37.2%.

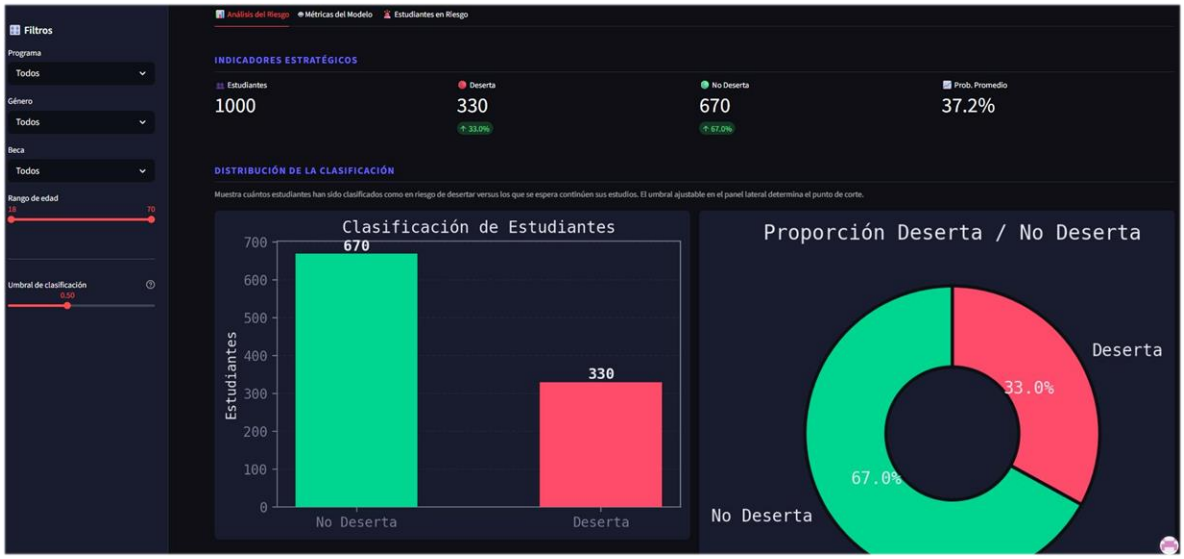


Figura 23. Predicciones realizadas por el sistema

6.3.2. Deserción por programa

La sección de deserción por programa académico presenta un gráfico de barras horizontales agrupadas que compara, para cada programa, el número de estudiantes clasificados como "Deserta" versus "No Deserta". Los programas se ordenan de menor a mayor número de desertores predichos, facilitando la identificación rápida de los programas con mayor concentración de riesgo institucional. Los códigos numéricos originales de cada programa fueron reemplazados por sus denominaciones completas para mejorar la interpretabilidad de la visualización.

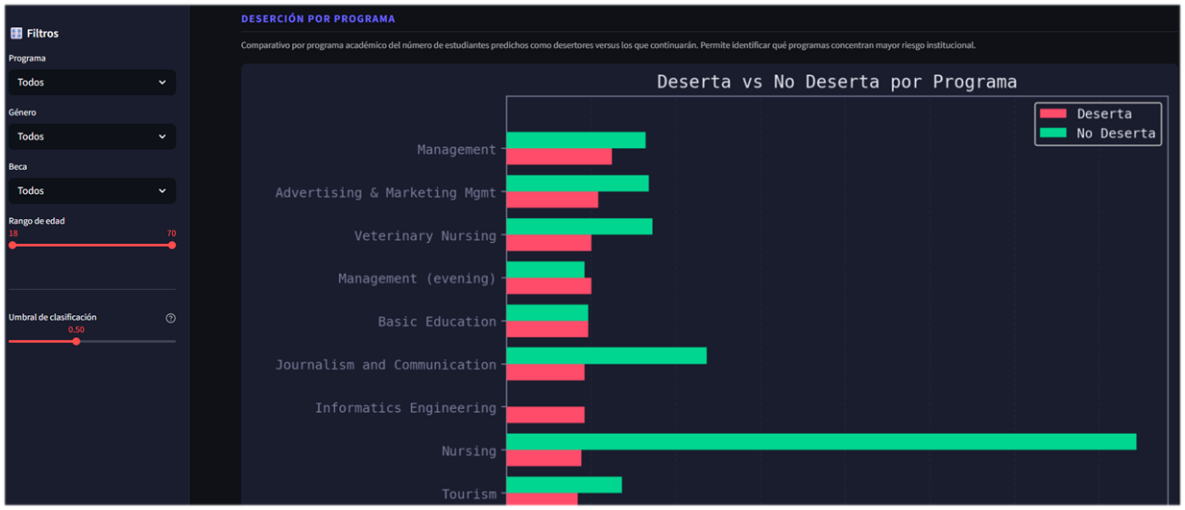


Figura 24. Predicciones por programa

6.3.3. Riesgo por variables demográficas

Esta sección presenta cuatro gráficas organizadas en dos filas. La primera fila incluye: (a) la probabilidad promedio de deserción diferenciada por género (Femenino/Masculino) y (b) la probabilidad promedio de deserción según la condición de beca (Con beca/Sin beca). En ambas gráficas se superpone una línea de referencia punteada que indica el umbral de clasificación activo, permitiendo evaluar visualmente qué grupos demográficos superan el punto de corte en promedio. La segunda fila presenta el análisis por edad: (a) un histograma de distribución de las edades de matrícula de la población cargada, y (b) un gráfico de barras que muestra la probabilidad promedio de deserción por rango etario, con codificación de color que destaca en rojo los rangos que superan el umbral de riesgo. Según lo observado en la Figura 25 el grupo de género masculino presenta una probabilidad promedio de deserción superior al umbral del 50%, mientras que los estudiantes con beca muestran una probabilidad considerablemente menor en comparación con aquellos sin beca, lo que evidencia el efecto protector del apoyo financiero sobre la permanencia estudiantil.

Figura 25. Riesgo por variables demográficas

La sección de análisis por edad comprende dos visualizaciones complementarias que permiten examinar la relación entre la edad de matrícula y el riesgo de deserción. La primera es un histograma que muestra la distribución de frecuencias de las edades de los estudiantes cargados al sistema, ofreciendo una visión general de la composición etaria de la población analizada. La segunda es un gráfico de barras que presenta la probabilidad promedio de deserción agrupada en cinco rangos etarios: 15-20, 20-25, 25-30, 30-40 y 40 años en adelante. Las barras se codifican en rojo cuando el promedio del rango supera el umbral de clasificación activo, y en verde cuando se mantiene por debajo de dicho umbral.

El gráfico de riesgo promedio por rango etario aporta información estratégica para el diseño de intervenciones diferenciadas. Los rangos etarios que superan el umbral de clasificación del 50% se destacan visualmente en rojo, señalando los grupos de edad que, en promedio, presentan mayor vulnerabilidad a la deserción. Esta visualización permite a los equipos institucionales orientar recursos de acompañamiento hacia los grupos de mayor edad, quienes frecuentemente enfrentan cargas laborales, familiares y económicas adicionales que aumentan su riesgo de abandono [47].

La combinación del histograma con el gráfico de riesgo por rango etario ofrece una perspectiva integral: no solo permite identificar qué grupos etarios son más vulnerables, sino también dimensionar el volumen de estudiantes que pertenecen a cada uno de ellos, facilitando la priorización de las acciones de intervención de acuerdo con el impacto potencial sobre la tasa de retención institucional.

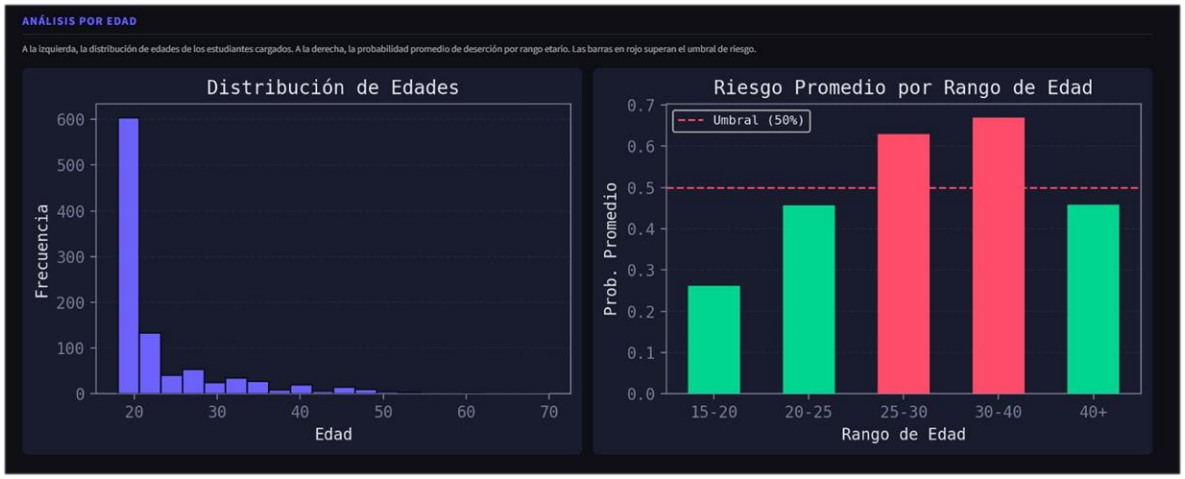


Figura 26. Análisis por edad

6.4. MODULO 2: MÉTRICAS DEL MODELO

El segundo módulo expone las métricas de desempeño del modelo XGBoost sobre el conjunto de entrenamiento, junto con visualizaciones que permiten evaluar su capacidad discriminativa y el peso relativo de las variables predictoras. Dado que los datos cargados al dashboard corresponden a nuevos registros de producción sobre los cuales no se dispone de etiquetas reales, las métricas presentadas corresponden exclusivamente al proceso de entrenamiento documentado en el Capítulo II.

Los cinco indicadores de desempeño del modelo presentados en la parte superior del módulo son los siguientes: Accuracy de 0,9649, Precision de 0,9675, Recall de 0,9420, F1-Score de 0,9545 y ROC AUC de 0,9948. Estos valores evidencian un alto poder predictivo del modelo sobre los datos de entrenamiento, siendo el ROC AUC un indicador especialmente relevante ya que cuantifica la capacidad del modelo para discriminar entre las clases "Deserta" y "No Deserta" independientemente del umbral de clasificación [34].

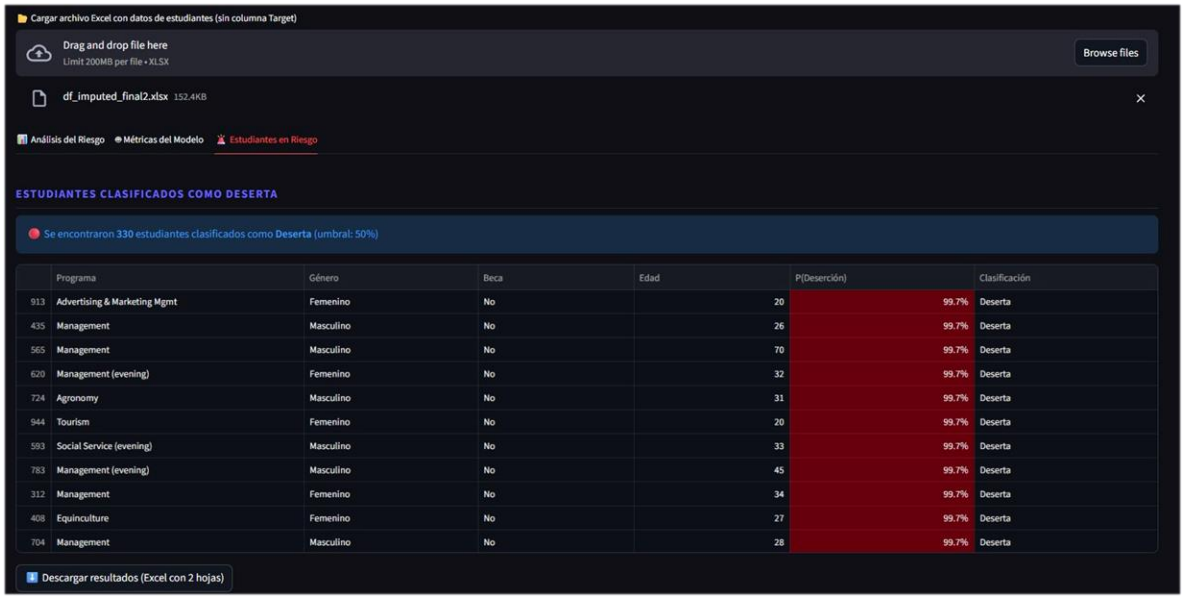
La Figura 27 muestra la curva ROC del conjunto de entrenamiento, cuya área bajo la curva (AUC = 0,9948) confirma la excelente capacidad discriminativa del modelo. Adicionalmente, el gráfico de importancia de variables (Feature Importance) revela que las unidades curriculares aprobadas en el segundo semestre constituyen el predictor de mayor relevancia, seguidas del estado de pago de la matrícula (Tuition fees up to date) y las unidades curriculares aprobadas en el primer semestre. Este hallazgo es consistente con la literatura que señala el rendimiento académico acumulado como el factor más determinante en la predicción de la deserción estudiantil [47].

Figura 27. Módulo de métricas del modelo

6.5. MODULO 3: ESTUDIANTES EN RIESGO

El tercer módulo constituye la interfaz operativa del sistema, orientada a la gestión individualizada del riesgo estudiantil. Presenta una tabla interactiva con el listado completo de los estudiantes clasificados como "Deserta", ordenados de mayor a menor probabilidad de deserción. Para cada registro se muestra el programa académico, género, condición de beca, edad al momento de matrícula, probabilidad de deserción estimada y clasificación final asignada.

La columna de probabilidad de deserción utiliza un gradiente de color rojo cuya intensidad es proporcional al valor de la probabilidad, permitiendo identificar visualmente los casos de mayor urgencia. Según lo presentado en la Figura 28, los registros en la parte superior de la tabla corresponden a estudiantes con probabilidades de deserción superiores al 99%, lo que los convierte en casos prioritarios para intervención institucional.



	Programa	Género	Beca	Edad	P(Deserción)	Clasificación
913	Advertising & Marketing Mgmt	Femenino	No	20	99.7%	Deserta
435	Management	Masculino	No	26	99.7%	Deserta
565	Management	Masculino	No	70	99.7%	Deserta
620	Management (evening)	Femenino	No	32	99.7%	Deserta
724	Agronomy	Masculino	No	31	99.7%	Deserta
944	Tourism	Femenino	No	20	99.7%	Deserta
593	Social Service (evening)	Masculino	No	33	99.7%	Deserta
783	Management (evening)	Masculino	No	45	99.7%	Deserta
312	Management	Femenino	No	34	99.7%	Deserta
408	Equinculture	Femenino	No	27	99.7%	Deserta
704	Management	Masculino	No	28	99.7%	Deserta

Figura 28. Módulo de estudiantes en riesgo

Al pie del módulo se dispone de un botón de descarga que genera un archivo Excel con dos hojas: (1) "Todos los estudiantes", con el resultado completo de la predicción para todos los registros cargados, y (2) "Deserta", con el subconjunto filtrado de estudiantes en riesgo. Este archivo constituye un insumo directo para los equipos de bienestar institucional y coordinación académica, facilitando la implementación de estrategias de intervención focalizadas.

6.6. DESPLIEGUE DEL PROTOTIPO

El prototipo fue desplegado en Streamlit Community Cloud, plataforma de alojamiento gratuito provista por Snowflake Inc. para aplicaciones desarrolladas con Streamlit [48]. El código fuente de la aplicación y los artefactos del modelo se alojaron en un repositorio público de GitHub, desde el cual Streamlit Community Cloud realiza el despliegue automático de la aplicación. Este esquema garantiza la disponibilidad permanente del sistema sin necesidad de mantener un servidor local activo. El repositorio del proyecto contiene los siguientes archivos: app.py (código de la aplicación), modelo_desercion.pkl, columnas_modelo.json, metricas_train.json, roc_train.json, feature_importance.csv, Variables_y_ejemplo.xlsx, y requirements.txt. Este último especifica las dependencias de Python necesarias para que Streamlit Community Cloud instale el entorno de ejecución correctamente. El prototipo se encuentra disponible de forma permanente en el siguiente enlace: https://deserci-n-estudiantil-n6jwnc6kki6wjpec2j6tqd.streamlit.app/

7. CONCLUSIONES Y TRABAJOS FUTUROS

7.1. CONCLUSIONES

El presente estudio permitió demostrar que la construcción de modelos predictivos de deserción estudiantil mediante técnicas de aprendizaje automático supervisado constituye una estrategia metodológicamente sólida y aplicable en contextos reales de educación superior. La adopción de la metodología CRISP-DM facilitó un abordaje estructurado del problema, donde la fase de preparación de datos se consolidó como un componente crítico para garantizar la calidad del modelado y la validez de los resultados obtenidos.

En comparación con los estudios revisados en el estado del arte, varios de los cuales evalúan uno o dos algoritmos, el presente trabajo implementó y comparó sistemáticamente cinco modelos de diferente complejidad estructural, integrando además un prototipo funcional desplegado en la nube, lo que constituye un aporte importante al estudio de la deserción.

La evaluación comparativa de los algoritmos evidenció que los modelos de mayor complejidad estructural presentan un mejor desempeño en la identificación de patrones asociados a la deserción. En este sentido, XGBoost fue seleccionado como el modelo más adecuado, al alcanzar el mayor Recall (0.9085) y una elevada capacidad discriminativa (ROC-AUC = 0.9728), lo que resulta especialmente relevante en contextos donde la prioridad institucional es minimizar la no detección de estudiantes en riesgo real de abandono.

El análisis de importancia de variables confirmó que el rendimiento académico temprano y el cumplimiento de obligaciones financieras son los principales factores asociados al abandono estudiantil, hallazgo consistente en los cinco modelos evaluados. Esta consistencia inter-modelo fortalece la validez del resultado y lo convierte en un insumo directo para el diseño de estrategias institucionales de intervención centradas en el monitoreo del desempeño académico desde el primer semestre y el acompañamiento financiero oportuno, reconociendo el carácter multifactorial del fenómeno.

Desde una perspectiva aplicada, el desarrollo del prototipo en Streamlit permitió materializar los resultados del modelo en una herramienta funcional, accesible y orientada a la toma de decisiones institucionales. Si bien en la evaluación comparativa de los cinco modelos se empleó el umbral estándar de 0.5 como punto de corte para la clasificación binaria, garantizando condiciones equitativas de comparación, el prototipo desarrollado incorpora un control deslizante que permite a los equipos institucionales ajustar dicho umbral según sus recursos de intervención disponibles y su tolerancia al riesgo. De este modo, instituciones con mayor capacidad de acompañamiento pueden reducir el umbral para ampliar la cobertura del sistema de alerta, mientras que aquellas con recursos más limitados pueden elevarlo para concentrar las intervenciones en los casos de mayor certeza.

Adicionalmente, la posibilidad de segmentar poblaciones específicas otorga flexibilidad operativa adicional, facilitando la implementación de estrategias de intervención diferenciadas según el perfil de riesgo de cada subpoblación estudiantil.

Finalmente, se reconoce que una limitación del presente estudio es el uso de datos provenientes de una institución europea, lo que implica que los patrones identificados podrían no transferirse directamente al contexto colombiano o latinoamericano sin una validación previa. En particular, la variable Tuition fees up to date no tiene un equivalente directo en el sistema colombiano, donde el acceso a financiamiento estudiantil opera mediante mecanismos distintos (créditos del ICETEX y esquemas de pago propios de cada institución), por lo que su distribución y su relación con la deserción probablemente difieran de las observadas en el contexto portugués. De manera similar, Scholarship holder refleja un esquema de becas europeo que no captura la diversidad de apoyos socioeconómicos existentes en Colombia (estratificación socioeconómica, subsidios de sostenimiento, becas municipales o departamentales), mientras que los indicadores macroeconómicos contextuales (Unemployment rate, Inflation rate, GDP), al corresponder a la economía portuguesa durante el periodo de estudio, no son aplicables directamente a un análisis colombiano y deberían sustituirse por sus equivalentes nacionales o regionales en una futura validación. Estas diferencias estructurales sugieren que, si bien la metodología y el flujo de trabajo propuestos son transferibles, las variables específicas y sus relaciones con la deserción deben recalibrarse con datos locales antes de su implementación institucional.

No obstante, los resultados evidencian que los modelos predictivos basados en aprendizaje automático pueden integrarse eficazmente en entornos institucionales reales, siempre que se acompañen de procesos organizacionales adecuados. El valor del sistema desarrollado radica no únicamente en su desempeño estadístico, sino en su capacidad para apoyar decisiones oportunas que contribuyan a la retención estudiantil y al fortalecimiento de la calidad educativa.

7.2. TRABAJOS FUTUROS

Una línea prioritaria de trabajo futuro es la validación del modelo XGBoost en instituciones de educación superior colombianas, utilizando datos reales de estudiantes locales. El conjunto de datos empleado en el presente estudio proviene del contexto europeo, por lo que resulta necesario evaluar si el modelo mantiene su desempeño cuando se aplica a contextos sociodemográficos, académicos e institucionales propios de la realidad colombiana y latinoamericana. Esta validación permitiría identificar qué variables requieren adaptación y si la estructura multifactorial del fenómeno se preserva en contextos distintos. Adicionalmente, se propone analizar la estabilidad temporal del modelo, verificando si las relaciones entre variables predictoras y deserción se mantienen constantes a lo largo de

varios semestres académicos, o si el modelo requiere reentrenamiento periódico para adaptarse a cambios en los patrones de abandono estudiantil.

Asimismo, se recomienda profundizar en el proceso de ingeniería de características explorando variables derivadas no incluidas en el presente estudio, tales como la razón entre la edad del estudiante y la edad promedio de su cohorte o programa académico, y la proporción de unidades curriculares aprobadas respecto a las matriculadas en cada periodo. Este tipo de variables relativas podría capturar de manera más precisa el nivel de ajuste del estudiante a su entorno académico, complementando los indicadores absolutos utilizados en los modelos actuales.

Los modelos implementados se construyeron a partir de variables académicas, socioeconómicas y personales disponibles en registros institucionales tradicionales. Una línea de trabajo futuro con alto potencial es la incorporación de datos provenientes de fuentes complementarias, tales como registros de participación en plataformas virtuales de aprendizaje, patrones de asistencia, frecuencia de uso de servicios de bienestar universitario y registros de asesoramiento académico. Estas variables podrían capturar señales tempranas de desvinculación que no están reflejadas en indicadores académicos formales, mejorando la anticipación del sistema. En este mismo sentido, la integración del sistema predictivo con plataformas académicas existentes mediante conexiones API permitiría automatizar el ingreso de datos, mantener predicciones actualizadas en tiempo real y eliminar procesos manuales de carga de información, facilitando la adopción institucional del sistema.

Se recomienda realizar auditorías detalladas del modelo para identificar y mitigar posibles sesgos en las predicciones según grupo demográfico. Es fundamental garantizar que el sistema trate equitativamente a todos los estudiantes y no perpetúe ni amplifique desigualdades preexistentes, lo cual es especialmente relevante en contextos de alta diversidad poblacional como el colombiano. Paralelamente, se propone desarrollar metodologías para determinar umbrales óptimos de clasificación que consideren no solo métricas estadísticas, sino también los costos e impactos reales de los falsos positivos y falsos negativos en contextos institucionales específicos. En particular, ajustar el umbral de clasificación de la red neuronal podría mejorar su Recall sin sacrificar significativamente su elevada precisión, convirtiéndola en una alternativa competitiva para instituciones con mayor capacidad de intervención.

Asimismo, se recomienda incorporar pruebas de significancia estadística pareadas, tales como el test de McNemar para comparar las tasas de error de clasificación entre pares de modelos, o el test de DeLong para evaluar diferencias estadísticamente significativas entre

curvas ROC, de manera que la selección del modelo final no se sustente únicamente en la comparación de estimaciones puntuales de las métricas de desempeño. Para su correcta implementación, se sugiere homogeneizar la semilla de partición entrenamiento/prueba entre todos los modelos evaluados, garantizando que las comparaciones pareadas se realicen sobre exactamente el mismo conjunto de observaciones.

Una extensión natural del presente trabajo es adaptar la metodología para predecir no solo deserción, sino también otras condiciones académicas críticas como riesgo de bajo rendimiento, rezago curricular o potencial de excelencia académica, construyendo así un sistema integral de identificación estudiantil que apoye tanto estrategias de retención como programas de acompañamiento diferenciado. En esta misma dirección, se propone explorar el desarrollo de modelos bajo esquemas de aprendizaje federado, donde múltiples instituciones contribuyan al entrenamiento de un modelo compartido sin necesidad de centralizar datos sensibles de sus estudiantes, aprovechando la diversidad de contextos institucionales para mejorar la robustez del modelo y preservando al mismo tiempo la privacidad de la información conforme a los marcos normativos de protección de datos personales.

Finalmente, los estudios de deserción suelen centrarse en evaluar el desempeño de los modelos predictivos sin medir el impacto real de las intervenciones derivadas de sus predicciones. Por ello, se recomienda diseñar estudios experimentales o cuasiexperimentales que cuantifiquen el efecto de las estrategias de retención implementadas a partir de las alertas generadas por el modelo, evaluando indicadores como la tasa de permanencia estudiantil y el rendimiento académico de los estudiantes intervenidos. Solo este tipo de evaluación permite establecer el valor institucional efectivo del sistema predictivo más allá de sus métricas de desempeño, vinculando la capacidad de predicción con resultados concretos en la trayectoria estudiantil.

8. REFERENCIAS BIBLIOGRÁFICAS

[1] M. I. Moreno Duarte, Análisis de la deserción universitaria en Colombia a partir de la revisión bibliográfica 2015–2021, Bogotá, Colombia, 2022.

[2] C. Romero and S. Ventura, "Educational data mining: A survey of the state of the art," IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews, vol. 40, no. 6, pp. 601–618, Nov. 2010.

[3] A. Namoun and A. Alshanqiti, "Predicting student performance using data mining and learning analytics techniques: A systematic literature review," Applied Sciences, vol. 11, no. 1, p. 237, 2020.

[4] V. Tinto, Completing College: Rethinking Institutional Action. Chicago, IL, USA: University of Chicago Press, 2012.

[5] D. Gutiérrez, J. F. Vélez Díaz and J. López, "Indicadores de deserción universitaria y factores asociados," EducaT: Educación Virtual, Innovación y Tecnologías, vol. 2, no. 1, pp. 15–26, 2021.

[6] J. Berens, K. Schneider, S. Gortz, S. Oster and J. Burghoff, "Detección temprana de estudiantes en riesgo: predicción de la deserción escolar utilizando datos administrativos y métodos de aprendizaje automático," Rev. Minería de Datos Educativos, vol. 11, no. 3, pp. 1–41, 2019. [En línea]. Disponible en: https://doi.org/10.5281/zenodo.3594771

[7] D. Delen, "A comparative analysis of machine learning techniques for student retention management," Decision Support Systems, vol. 49, no. 4, pp. 498–506, 2010.

[8] K. L. García Mendoza, J. Rodríguez Heras, Y. Ferrer Mendoza and Z. C. Fragozo Torres, "Deserción estudiantil en Colombia: un análisis de los factores predictores y sus indicadores," Encuentros: Revista de Ciencias Humanas, Teoría Social y Pensamiento Crítico, Extra, pp. 297–310, 2022. doi: 10.5281/zenodo.6551173.

[9] J. C. González Sánchez and M. J. Peñaloza Pérez, "Identificación y predicción de estudiantes en riesgo de deserción académica por medio de modelos basados en Machine Learning," 2021.

[10] M. Bassi, S. Busso, S. Urzúa and J. Vargas, Desconectados: Habilidades, educación y empleo en América Latina. Washington, D.C.: Banco Interamericano de Desarrollo, 2012.

[11] E. Castaño, S. Gallón, K. Gómez y J. Vásquez, "Deserción estudiantil universitaria: una aplicación de modelos de duración," Lecturas de Economía, vol. 60, no. 60, pp. 39–65, 2004.

[12] Ministerio de Educación Nacional de Colombia, "Informe de deserción estudiantil en Colombia," 2022. [En línea]. Disponible en: https://www.mineducacion.gov.co. [Accedido: 10 mayo 2025].

[13] Departamento Administrativo Nacional de Estadística (DANE), Impacto de factores laborales en la educación superior, 2020. [En línea]. Disponible en: https://www.dane.gov.co/. [Accedido: 10 mayo 2025].

[14] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York: Springer, 2009.

[15] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. 2nd ed. Sebastopol, CA: O'Reilly Media, 2019.

[16] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression. 3rd ed. Hoboken, NJ: John Wiley & Sons, 2013.

[17] M. Kuhn and K. Johnson, Applied Predictive Modeling. New York: Springer, 2013.

[18] T. M. Mitchell, Machine Learning. New York, NY, USA: McGraw-Hill, 1997.

[19] J. R. Quinlan, "Improved use of continuous attributes in C4.5," Journal of Artificial Intelligence Research, vol. 4, pp. 77–90, 1996.

[20] J. Han, M. Kamber y J. Pei, Data Mining: Concepts and Techniques. Amsterdam, Países Bajos: Elsevier, 2011.

[21] C. Arana, "Modelos de aprendizaje automático mediante árboles de decisión," Documento de Trabajo, no. 778, 2021.

[22] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[23] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning. New York: Springer, 2013.

[24] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.

[25] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, Aug. 2016, pp. 785–794.

[26] K. Pandey and S. Srivastava, "Student performance prediction using machine learning over educational data mining," Journal of Education and Learning, vol. 15, no. 4, pp. 541–552, 2021.

[27] J. D. Kelleher, B. Mac Namee, and A. D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies. 2nd ed. Cambridge: MIT Press, 2020.

[28] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," The Bulletin of Mathematical Biophysics, vol. 5, no. 4, pp. 115–133, 1943.

[29] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015.

[30] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," Nature, vol. 323, no. 6088, pp. 533–536, Oct. 1986.

[31] I. H. Witten, E. Frank, M. A. Hall y C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, 4a ed. Burlington, MA, USA: Morgan Kaufmann, 2016. doi: 10.1016/c2009-0-19715-5.

[32] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.

[33] D. J. Hand, "Measuring classifier performance: A coherent alternative to the area under the ROC curve," Machine Learning, vol. 77, pp. 103–123, 2009.

[34] J. Huang y C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," IEEE Trans. Knowl. Data Eng., vol. 17, no. 3, pp. 299–310, 2005.

[35] M. V. Diaz Lema, M. Cannistrà, C. van Klaveren, T. Agasisti e I. Cornelisz, "Predicting dropout in Higher Education across borders," Studies in Higher Education, vol. 49, no. 1, pp. 141–156, 2024. doi: 10.1080/03075079.2023.2224818.

[36] O. Sifuentes, "Modelos predictivos de la deserción estudiantil en una universidad privada peruana," Industrial Data, vol. 21, no. 2, pp. 47–62, 2018. [En línea]. Disponible en: https://www.redalyc.org/journal/816/81658967008/html/. [Accedido: 10 mayo 2025].

[37] A. G. Hernandez Gonzalez, R. A. Melendez Armenta, L. A. Morales Rosales, A. Garcia Barrientos, J. L. Tecpanecatl Xihuitl y I. Algredo, "Comparative Study of Algorithms to Predict the Desertion in the Students at the ITSM-Mexico," IEEE Latin America Transactions, vol. 14, no. 11, pp. 4573–4578, 2016. doi: 10.1109/TLA.2016.7795831.

[38] B. Cuji, W. Gavilanes y R. Sanchez, "Modelo predictivo de deserción estudiantil basado en árboles de decisión," Espacios, vol. 38, no. 55, p. 17, 2017. [En línea]. Disponible en: http://www.revistaespacios.com/a17v38n55/a17v38n55p17.pdf. [Accedido: 25 May. 2025].

[39] A. Camargo, Modelo para la predicción de la deserción de estudiantes de pregrado, basado en técnicas de minería de datos. Corporación Universitaria de la Costa, 2020. [En línea]. Disponible en: https://repositorio.cuc.edu.co/handle/11323/7077. [Accedido: 10 mayo 2025].

[40] Universidad de los Andes, Diseño de un modelo predictivo sobre la deserción de un estudiante de ingeniería eléctrica y electrónica de la Universidad de los Andes utilizando técnicas de Machine Learning. Bogotá, Colombia, 2020.

[41] J. Vásquez, "Modelo predictivo para estimar la deserción de estudiantes en una institución de educación superior," Rev. Educ. y Desarrollo, vol. 45, pp. 25–35, 2023. [En línea]. Disponible en: https://www.scielo.org.mx/scielo.php?pid=S1607-40412023000100113&script=sci_arttext [Accedido: 25 May. 2025].

[42] M. A. García, "Predicting student dropouts with machine learning: An empirical study," Technology in Society, vol. 76, 2024. [En línea]. Disponible en: https://www.sciencedirect.com/science/article/pii/S0160791X24000228 [Accedido: 25 May. 2025].

[43] F. R. Munizaga Mellado, M. B. Cifuentes Orellana y A. J. Beltrán Gabrie, "Retención y deserción estudiantil en la educación superior en América Latina y el Caribe: una revisión sistemática," EPAA, vol. 26, p. 61, mayo de 2018.

[44] V. Realinho, M. Vieira Martins, J. Machado, y L. Baptista, "Predict Students' Dropout and Academic Success," UCI Machine Learning Repository, 2021. [En línea]. Disponible en: https://doi.org/10.24432/C5MC89

[45] Streamlit Inc., "Streamlit — A faster way to build and share data apps," 2023. [Online]. Available: https://streamlit.io

[46] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[47] V. Tinto, "Dropout from Higher Education: A Theoretical Synthesis of Recent Research," Review of Educational Research, vol. 45, no. 1, pp. 89–125, 1975.

[48] Snowflake Inc., "Streamlit Community Cloud," 2023. [Online]. Available: https://share.streamlit.io

ANEXOS

Anexo 1. Conjunto de datos. Enlace: https://doi.org/10.24432/C5MC89

Anexo 2. Prototipo de visualización. Enlace: https://deserci-n-estudiantil-n6jwnc6kki6wjpec2j6tqd.streamlit.app/