Desarrollo de modelo clasificación para determinar el riesgo de deserción estudiantil de los beneficiarios de pregrado en una entidad que ofrece créditos estudiantiles en Medellín

Genaro Alfonso Aristizabal Echeverri - Euler Leonardo Cuarán Rosero

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor

Gabriel Darío Uribe Guerra, Ms.C Matemáticas

Universidad de Antioquia

Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2025

Cita	(Aristizabal Echeverri & Cuarán Rosero, 2025)
Referencia **Estilo APA 7 (2020)**	Aristizabal Echeverri, G. A., & Cuarán Rosero, E. L. (2025). Desarrollo de modelo de clasificación para determinar el riesgo de deserción estudiantil de los beneficiarios de pregrado en una entidad que ofrece créditos estudiantiles en Medellín. Universidad de Antioquia, Medellín, Colombia.

Especialización en Analítica y Ciencia de Datos, Cohorte VIII.

Centro de Investigación Ambientales y de Ingeniería (CIA).

Dedicatoria

Dedicamos este trabajo a nuestras familias, por su apoyo incondicional, su paciencia y la confianza que siempre depositaron en nosotros.

Agradecimientos

Agradecemos sinceramente a nuestras familias y amistades por estar presentes en cada etapa de este proceso. A nuestros docentes y compañeros, por compartir sus conocimientos, experiencias y motivación. Y a la entidad gubernamental que nos proporcionó la información, por su amabilidad, transparencia y, sobre todo, por su buena disposición y confianza en el proyecto.

Tabla de contenido

Lista de Figuras

Lista de Tablas

Siglas, acrónimos y abreviaturas

IES	Institución de educación Superior
SNIES	Sistema Nacional de Información de la educación superior
PUAP	Programa único de Acceso y Permanencia
SPADIES	Sistema para la prevención y análisis de la deserción en la educación superior
ONU	Organización de las Naciones Unidas
PP	Presupuesto Participativo
RO	Recurso Ordinario
SISBEN	Sistema de Identificación de Potenciales Beneficiarios de Programas Sociales.
ML:	Machine Learning
NB	Naive Bayes
RF	Random Forest
CB	CatBoost
Esp.	Especialista
MSc	Magister Scientiae
Párr.	Párrafo
UdeA	Universidad de Antioquia

Resumen - En Colombia, diversas entidades, tanto públicas como privadas, comprometidas con el fortalecimiento de la educación superior, han incentivado programas de créditos condonables para facilitar el acceso y la permanencia de los estudiantes. En Medellín, una entidad pública que implementa este tipo de apoyos ha buscado fortalecer sus estrategias institucionales para anticipar los casos de deserción estudiantil. Por ello, se creó un modelo de clasificación que utiliza técnicas de aprendizaje automático, con el objetivo de identificar a los beneficiarios activos que puedan ser clasificados como posibles desertores a lo largo de su trayectoria académica en la entidad. Para el desarrollo del modelo, se utilizó un registro histórico por estudiante, abarcando el periodo 2019–2025. A partir de este conjunto de datos, se seleccionaron once variables, elegidas tras realizar análisis descriptivos. Estas variables incluían información académica, transaccional, socioeconómica y sociodemográfica. El proceso metodológico abarcó etapas de limpieza, transformación de datos y la comparación de varios algoritmos de clasificación. En conjunto, el modelo sirvió como herramienta de apoyo al área de permanencia de la entidad, para la toma de decisiones encaminadas a acompañar a los estudiantes que tienen riesgo de desertar del programa.

Keywords— deserción estudiantil, créditos condonables, aprendizaje automático, CatBoost Classifier, Medellín.

1. Introducción

1.1. Planteamiento del Problema

En la educación superior se abarcan todas las modalidades de enseñanza (profesional, técnica, artística, pedagógica, etc.) impartida por universidades, institutos tecnológicos, escuelas normales, etc. (Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura [UNESCO], 2020, p.14). Estas instituciones están dirigidas a personas que han completado la educación secundaria y tienen como propósito la obtención de un título, grado, certificado o diploma en el ámbito de la educación superior. El acceso a la educación superior ha aumentado en todas las regiones del mundo durante las últimas dos décadas (UNESCO, 2020, p.24). Sin embargo, la deserción postsecundaria es uno de los principales desafíos a los que se enfrentan los estudiantes que aspiran a culminar sus estudios universitarios.

En Colombia, según lo especifica el SPADIES (sistema para la prevención y análisis de la deserción en la educación superior) un estudiante se clasifica como desertor del sistema de educación superior cuando no se ha matriculado por dos o más periodos consecutivos en algún programa académico (Ministerio de Educación Nacional de Colombia, 2023). Teniendo esto en cuenta y basándose en datos del SNIES (Sistema Nacional de Información de la educación superior) y el LEE (Laboratorio de Economía y Educación de la Universidad Javeriana) se estima que, entre el año 2000 y 2021, la tasa de deserción se ubicó en un 11% por periodo académico, lo que significa que en promedio cada semestre, 1 de cada 10 estudiantes a nivel nacional desertan de sus estudios postsecundarios ya sea de manera temporal o permanente (Ministerio de Educación Nacional de Colombia, 2023).

La deserción estudiantil en la educación superior representa un desafío crítico para las instituciones educativas, afectando tanto los indicadores de calidad como la gestión de recursos

económicos, especialmente en programas de apoyo como los de las entidades que promueven el acceso a la educación superior (UNESCO, 2020). La predicción temprana de la deserción de estudiantes de educación superior permite implementar estrategias preventivas más efectivas (Kang & Wang, 2018).

Diversos estudios han demostrado la utilidad de técnicas de aprendizaje automático para abordar el problema de la deserción estudiantil en la educación postsecundaria. Kang y Wang desarrollaron un modelo de regresión logística basado en datos institucionales de la Universidad de Lamar, en Texas, con el objetivo de identificar, antes del inicio de un nuevo periodo académico, a los estudiantes con mayor probabilidad de desertar; su modelo alcanzó una precisión del 81,8% (Kang & Wang, 2018). Albán y Mauricio (2018) aplicaron árboles de decisión a partir de encuestas realizadas en universidades ecuatorianas, logrando una precisión del 97,95% en la detección temprana de estudiantes en riesgo de deserción. Vilora, Pineda y Varela (2019) utilizaron un clasificador bayesiano simple con datos de la Universidad del Zulia y la Universidad Rafael Belloso Chacín, en Venezuela, obteniendo una tasa de detección del 33,35%. Guzmán Castillo et al. (2022) desarrollaron un sistema predictivo con información proveniente de universidades del Caribe colombiano, integrando variables sociodemográficas, académicas y de pruebas estandarizadas como el ICFES. Aulck et al. (2024) llevaron a cabo un estudio en la Universidad de Washington, en Seattle (Estados Unidos), donde demostraron la efectividad de modelos de predicción basados en datos institucionales. En Colombia, Quintero (2022) implementó modelos basados en redes neuronales y XGBoost utilizando información de la Facultad de Ingeniería de la Universidad de Antioquia, alcanzando una precisión del 74,91%, con mejor desempeño en estudiantes que ya habían cursado varios semestres.

A pesar de los avances metodológicos logrados por la literatura internacional y algunos estudios nacionales, la aplicación de estos enfoques en contextos locales específicos sigue siendo limitada. Tal es el caso de Medellín, donde las problemáticas asociadas al acceso y permanencia en la educación postsecundaria persisten. Para afrontar, el distrito cuenta con una entidad que lidera programas de apoyo financiero mediante créditos condonables, dirigidos a personas de bajos recursos. Estos créditos pueden ser condonados total o parcialmente si se cumplen requisitos como rendimiento académico y servicio social. Sin embargo, aún no se han implementado mecanismos predictivos ajustados a las características sociales y administrativas de estos programas, lo que abre la posibilidad de explorar soluciones basadas en modelos de clasificación automatizada. Aunque existen estudios recientes en Medellín que aplican modelos predictivos de deserción, como el de Quintero (2022) en la Universidad de Antioquia, no se han identificado investigaciones centradas exclusivamente en beneficiarios de programas de créditos condonables distritales. Esta población presenta características socioeconómicas y condiciones particulares, que justifican la necesidad de enfoques específicos para su análisis.

En este contexto, la aplicación de la ciencia de datos y el aprendizaje automático se presenta como una alternativa para identificar patrones asociados a la deserción estudiantil. A diferencia de los métodos cualitativos o enfoques tradicionales, los modelos de clasificación automatizada permiten analizar simultáneamente grandes volúmenes de datos socioeconómicos, académicos

y demográficos, facilitando la identificación de patrones complejos y la generación de alertas tempranas sobre el riesgo de deserción. Por tanto, en este estudio, se construyen modelos de clasificación orientados exclusivamente a identificar la deserción estudiantil postsecundaria, utilizando datos históricos de beneficiarios activos y no activos proporcionados por una entidad encargada de ofrecer créditos condonables educativos a estudiantes de diversas instituciones de educación técnica, tecnología y universitaria en Medellín.

1.2. Objetivo de estudio

Desarrollar un modelo de clasificación basado en técnicas de aprendizaje automático para identificar la deserción estudiantil en beneficiarios de una entidad que ofrece créditos condonables de pregrado para la educación postsecundaria en Medellín, utilizando variables socioeconómicas, sociodemográficas, académicas y financieras.

1.3. Metodología

Para este estudio se utilizó un conjunto de datos suministrado por el programa de créditos condonables para educación postsecundaria en Medellín, que abarca el periodo 2019–2025. Tras un proceso de limpieza, transformación y depuración, se consolidó una muestra final de 14.641 registros y 11 variables seleccionadas por su capacidad explicativa, relacionadas con aspectos socioeconómicos, sociodemográficos, académicos y financieros de los beneficiarios.

Con el fin de predecir la deserción estudiantil, se aplicaron tres algoritmos de clasificación supervisada: Naive Bayes, como línea base por su simplicidad y eficiencia; Random Forest, por su capacidad para manejar datos heterogéneos y evitar el sobreajuste; y CatBoost, por su efectividad al tratar variables categóricas y su alto rendimiento general. El desempeño de los modelos se evaluó mediante métricas estándar para problemas de clasificación binaria: exactitud, precisión, recall y F1-score.

2. Materiales y métodos

2.1. Descripción del conjunto de datos

Se cuenta con un registro histórico de los beneficiarios de una entidad que ofrece créditos condenables, abarcando el periodo comprendido entre los años 2019 y 2025, siguiendo un riguroso proceso de tratamiento y anonimización de datos personales, con forme a los lineamientos establecidos por la agencia de educación postsecundaria en el acuerdo de confidencialidad *habeas data.*

2.1.1. Fuente de los datos.

Los datos originales se extrajeron de las bases de datos SQL Server de los servidores de la agencia de educación postsecundaria, a través de consultas a las distintas tablas referentes a proyecciones financieras, formularios de inscripción, montos y estados históricos de renovación de los beneficiarios.

2.1.2. Tamaño y estructura

La base de datos original se exportó en formato Parquet, un tipo de almacenamiento columnar utilizado para el procesamiento eficiente de datos. El archivo posee un peso aproximado de 8 MB el cual contiene 101673 filas y 42 columnas.

2.1.3. Características principales

La base de datos original se compone de variables referentes a información académica, socioeconómica, sociodemográfica y transaccional de los beneficiarios. Donde se destacan las siguientes variables:

Tabla 1. *Variables y tipos de datos empleados en la predicción de deserción estudiantil.*

Nº	Nombre de la columna	Descripción de las variables	Tipo de dato
1	FONDO	Nombre de la fuente de financiación con la que se legalizo el crédito del beneficiario.	Categórica
2	SEMESTRES A FINANCIAR	Cantidad de semestres de la carrera que se le financian al beneficiario del crédito condonable.	Número entero
3	GIROS PENDIENTES	Cantidad de giros pendientes por desembolsar al beneficiario del crédito condonable.	Número entero
4	SEXO	Sexo del beneficiario.	Categórica
5	COMUNA DE RESIDENCIA	Comuna de residencia del beneficiario	Categórica
6	PUNTAJE SISBEN	Puntaje del SISBEN del beneficiario, se contemplan 5 grupos según su condición. A: Personas en pobreza extrema, B: Población en pobreza moderada, C: Población vulnerable, D: Población no vulnerable y NO TIENE / NO PRESENTA: Personas que no presentan puntaje de SISBEN.	Categórica
7	EDAD	Edad actual del beneficiario.	Número entero
8	Num_Sus_Temporales	Cantidad de periodos de renovación en los que el beneficiario suspendió temporalmente su crédito.	Número entero
9	Es_Desertor	Variable dicotómica que representa las siguientes categorías: 1: El beneficiario deserto del crédito y 0: El beneficiario no deserto del crédito.	Número entero
10	monto_total_girado	Corresponde con al dinero total girado al beneficiario hasta la fecha.	Número entero
11	VICTIMA DEL CONFLICTO ARMADO	Variable dicotómica que contiene las categorías SI: El beneficiario fue víctima del conflicto armado y NO: El beneficiario no fue víctima del conflicto armado.	Categórica

2.2. Procesamiento y limpieza de datos

El conjunto de datos utilizado en este estudio fue sometido a un proceso de limpieza y transformación para asegurar la calidad y la relevancia de las variables.

2.2.1. Identificación de valores nulos

Se identifico que las variables con registros faltantes eran: "TOTAL CREDITOS DE LA CARRERA", "SEMESTRE DE INGRESO", "ESTADO GENERAL" y "MOTIVO ESTADO". Para la variable "TOTAL CREDITOS DE LA CARRERA", se realizó un cruce con la base de datos de programas del SNIES (Sistema Nacional de Información de la Educación Superior) para obtener los créditos correspondientes a cada programa. En cuanto a "SEMESTRE DE INGRESO", se aplicó una imputación basada en la duración del programa y los semestres financiados, con el fin de completar los registros faltantes. Por otro lado, la variable "ESTADO GENERAL" fue recalculada a partir de la información de "ESTADO SEMESTRAL", mientras que la columna "MOTIVO ESTADO" fue ajustada para reflejar razones más claras relacionadas con el estado de los beneficiarios, como "Renovado", "No renovado" y "Culmino periodos pactados".

2.2.1. Eliminación de variables

Se eliminaron algunas variables que no aportaban un valor significativo al análisis, ya sea por su alta correlación entre ellas o porque no eran muy útiles para entender el fenómeno de la deserción. Primero, se descartaron variables transaccionales que estaban muy correlacionadas, como el monto de matrícula y el monto de sostenimiento, ya que ofrecían información redundante con respecto a la variable monto total girado. En segundo lugar, se excluyeron variables categóricas relacionadas con procesos académicos, como el programa académico, la institución de educación superior (IES), la modalidad de crédito y la cantidad de suspensiones especiales. Estas variables tenían un gran desbalance en sus categorías, muchas veces eran redundantes con otras ya incluidas en el modelo, y no ofrecían información relevante para explicar el fenómeno de la deserción estudiantil.

2.2.3. Transformación de variables

Se corrigieron inconsistencias en los registros de la comuna de residencia. Así mismo, se limpiaron y estandarizaron los valores de algunas variables categóricas, como el "ESTRATO", unificando categorías que aparecían con variaciones en la nomenclatura. Adicionalmente, se llevó a cabo la conversión de algunas variables numéricas y categóricas como la fecha de nacimiento y el "PUNTAJE SISBEN" que fue estandarizado según las categorías A, B, C y D actuales.

2.2.4. Condensación de registros

Puesto que la base de datos original contaba con varios registros históricos de cada beneficiario por periodo renovado, se condensó toda la información de tal modo que, resulto un registro único por cada beneficiario pasando de tener 101673 a 14641 registros.

2.2.5. Análisis descriptivo y selección de características

Se realizó un análisis descriptivo con estadísticos univariados, diagramas de dispersión y análisis de correlaciones para evaluar la distribución de las variables y su relación con el estado de deserción. A partir de este proceso exploratorio, se seleccionaron las 11 variables más representativas, de las cuales 5 son categóricas y 6 numéricas. Estas variables condensan las siguientes dimensiones:

- **Transaccional**: incluye las variables *monto_total_girado*, *giros_pendientes*, *semestres_a_financiar* y *fondo*, que reflejan aspectos operativos del proceso de financiación.
- **Socioeconómica**: conformada por *puntaje_SISBEN* y *víctima_del_conflicto_armado*, como indicadores del nivel de vulnerabilidad económica.
- **Sociodemográfica**: compuesta por las variables *sexo*, *edad* y *comuna_de_residencia*, que caracterizan la población beneficiaria.
- **Académica**: integrada por *num_suspensiones_temporales*, *es_desertor* y nuevamente *semestres_a_financiar*, en tanto que esta última también guarda relación con la trayectoria académica.

El conjunto de datos resultante, tras los procesos de depuración y tratamiento, consta de 14.641 registros y 11 variables.

2.3. Métodos de análisis y modelado

Este estudio utiliza modelos como Naive Bayes (con sus variantes GaussianNB, BernoulliNB y MultinomialNB), Random Forest y CatBoost Classifier para la identificación de la deserción estudiantil, que pertenece a un problema de clasificación binaria que determina si el estudiante abandonará o no sus estudios post secundarios. Los modelos seleccionados son reconocidos por su capacidad para manejar datos categóricos y numéricos, además de su eficacia en tareas de clasificación supervisada.

Para garantizar un enfoque riguroso, el conjunto de datos se dividió en un 80 % para entrenamiento y un 20 % para prueba. Adicionalmente, se aplicó una técnica denominada *undersampling* la cual reduce el número de observaciones de todas las clases excepto en la clase minoritaria, esto con el objetivo de mitigar el desbalanceo de las clases, específicamente en los modelos Random Forest y Naive Bayes. También, las variables numéricas fueron estandarizadas mediante normalización Z-score, transformando sus valores en función de la media y la desviación estándar, lo cual es necesario para evitar que las diferencias de escala influyan en el rendimiento de los modelos. En el caso de CatBoost, este tipo de preprocesamiento no fue necesario, dada su capacidad para manejar variables categóricas de forma nativa y su solvencia frente al desbalance de clases.

A continuación, se describen los modelos utilizados junto con su configuración y entrenamiento específico.

2.3.1. Naive Bayes

Es un clasificador probabilístico basado en el teorema de Bayes con fuertes suposiciones de independencia entre las características. El desacoplamiento de las distribuciones de variables condicionales de clase permite que cada distribución se calcule de forma independiente como una distribución unidimensional (Pérez, Castellanos, & Correal, 2018). Cada variante de Naive Bayes fue entrenada utilizando el método *fit* sobre el conjunto de entrenamiento preprocesado y evaluada a través de validación cruzada, con el fin de comparar su desempeño bajo distintos supuestos probabilísticos.

2.3.2. Random Forest

Se compone de múltiples árboles de decisión construidos a partir de muestras obtenidas con reemplazo (bootstrapping) y submuestreo aleatorio de características, lo que reduce la correlación entre árboles y mejora la generalización del modelo (Breiman, 2011). En problemas de clasificación, la predicción final se obtiene por voto mayoritario.

Para este estudio, se utilizó la clase *RandomForestClassifier* de Scikit-learn, configurado inicialmente con n_estimators=100 y random_state=42 para garantizar reproducibilidad (Martínez & Castillo, 2024). Este conjunto inicial de hiperparámetros se ajustó mediante una búsqueda de hiperparámetros con validación cruzada usando RandomizedSearchCV. Los parámetros ajustados incluyeron el número de árboles (n_estimators), la profundidad máxima (max_depth), el número mínimo de muestras para dividir un nodo (min_samples_split), y el número mínimo de muestras por hoja (min_samples_leaf).

El entrenamiento del modelo se realizó con el método fit sobre el conjunto de entrenamiento, incluyendo validación cruzada para controlar el sobreajuste y evaluar la estabilidad del modelo.

2.3.3. CatBoost Classifier

Es un algoritmo de aprendizaje automático basado en gradient boosting que utiliza árboles de decisión binarios como predictores base. Se destaca por su eficiencia al procesar variables categóricas, ya que las transforma internamente sin necesidad de codificación previa como one-hot (Ibrahim, Ridwan, Muhammed, Abdulaziz, & Saheed, 2020). CatBoost construye los modelos de forma iterativa, corrigiendo errores en cada paso y minimizando la función de pérdida esperada. Su mecanismo de ordenamiento por permutación reduce el riesgo de sobreajuste, lo que lo convierte en una opción precisa y robusta para problemas de clasificación, como la predicción de la deserción estudiantil (Ibrahim, Ridwan, Muhammed, Abdulaziz, & Saheed, 2020).

Para este estudio, se implementó mediante la clase *CatBoostClassifier* de la biblioteca CatBoost, configurado inicialmente con iterations=1000, learning_rate=0.05, depth=8 y random_state=42 para garantizar la reproducibilidad (Sairete, y otros, 2021). Estos hiperparámetros fueron seleccionados considerando un balance entre rendimiento, capacidad de generalización y tiempo de entrenamiento. Posteriormente, se entrenaron múltiples modelos variando los hiperparámetros con el objetivo de minimizar el sobreajuste, monitoreando la

métrica de pérdida logarítmica (logloss) tanto en el conjunto de entrenamiento como en el de validación. La selección del modelo final se basó en la reducción del gap de logloss entre entrenamiento y validación, priorizando configuraciones que mantuvieran un buen equilibrio entre ajuste y generalización.

2.4 Métricas utilizadas en la evaluación de los modelos

Durante la evaluación, se emplearon métricas detalladas como la matriz de confusión, curva ROC, precisión-recall y análisis o curvas de calibración, lo que permitió medir tanto la capacidad del modelo para discriminar entre estudiantes que desertan y los que no, como la calidad y calibración de las probabilidades estimadas.

Tabla 2. *Parámetros de los modelos de aprendizaje automático empleados en la clasificación de la deserción estudiantil.*

Algoritmo	Parámetro	Descripción	Valor
CatBoost	learning_rate	Tasa de aprendizaje	0.01
	iterations	Número máximo de iteraciones (árboles)	500
	depth	Profundidad máxima de los árboles	7
	l2_leaf_reg	Parámetro de regularización L2 en las hojas	6
	eval_metric	Métrica usada para evaluación durante entrenamiento	AUC
	random_seed	Semilla para reproducibilidad	42
	early_stopping_rounds	Número de iteraciones para detener temprano si no hay mejora	30
Random Forest	criterion	Función para medir calidad de una división	gini
	max_depth	Profundidad máxima del árbol	16
	max_features	Número máximo de características consideradas en cada división	11
	n_estimators	Número de árboles en el bosque	100
	random_state	Semilla para reproducibilidad	123
Naive Bayes	var_smoothing (GaussianNB)	Parámetro para suavizado de varianza en GaussianNB	Default
	binarize (BernoulliNB)	Umbral para binarizar características	0.0
	alpha (MultinomialNB)	Parámetro de suavizado Laplace	1.0

3. Resultados y discusiones

En esta sección se presentan y analizan los resultados obtenidos a partir de la aplicación de cinco algoritmos de aprendizaje automático para clasificar la deserción estudiantil. Los modelos evaluados fueron tres variantes de Naive Bayes (GaussianNB, BernoulliNB y MultinomialNB), junto con Random Forest y CatBoost Classifier. La variable objetivo se definió como binaria (1: desertor, 0: no desertor), y se emplearon métricas clásicas para tareas de clasificación: precisión, recall y F1-score, usadas en estudios recientes (Sairete, y otros, 2021), evaluadas por clase. Además, se calculó la curva ROC con su AUC, la curva de recall, y se graficaron las variables con mayor importancia para cada modelo.

3.1. Comparación de modelos

En la Tabla 3, se presentan los indicadores de rendimiento más representativos: precisión, recall y F1-score, para las clases 0 y 1.

Tabla 3. *Resultados de precisión, recall y F1-score por clase para los modelos de clasificación de deserción estudiantil.*

Modelo	Clase	Precision	Recall	F1-score
GaussianNB	0	0.98	0.85	0.91
	1	0.48	0.90	0.62
BernoulliNB	0	0.95	0.70	0.81
	1	0.28	0.78	0.42
MultinomialNB	0	0.98	0.79	0.88
	1	0.40	0.90	0.56
RandomForest	0	0.99	0.94	0.96
	1	0.69	0.96	0.81
CatBoost	0	0.98	0.98	0.98
	1	0.86	0.86	0.86

Como se observa en la Tabla 3, el modelo CatBoost obtuvo el mejor desempeño para la clase 1 con un F1-score, precision y recall de 0.86. Esto indica que CatBoost fue especialmente eficaz para identificar correctamente tanto a los estudiantes que desertan como a los que no. En segundo lugar, el modelo Random Forest alcanzo un F1-score de 0.81 para la clase de estudiantes desertores, con un recall de 0.96 y una precisión de 0.69. Aunque este modelo logró un mayor recall (detección de desertores), lo hizo a costa de una menor precisión, lo que implica una mayor proporción de falsos positivos.

Los modelos basados en Naive Bayes no obtuvieron valores relevantes. Aunque GaussianNB logró un valor de F1-score de 0.91 para la clase no desertora, su rendimiento para la clase desertora fue menor (0.62). BernoulliNB fue el modelo con peor rendimiento general, con un F1-score de solo 0.42 para la clase 1.

En la **Figura *1*** se indican las curvas ROC de los mejores modelos, todos ellos presentan una tasa de verdaderos positivos (TPR) muy cercana a 1. Sin embargo, se observa como la curva

del modelo gausiano Naive Bayes presenta dificultades identificando a los falsos positivos, ya que tiende a clasificarlos erróneamente como no desertores. Por otro lado, se evidencia que los modelos generados por Random Forest y CatBoost logran capturar la gran mayoría de beneficiarios que efectivamente son desertores manteniendo muy baja la proporción de falsos positivos.

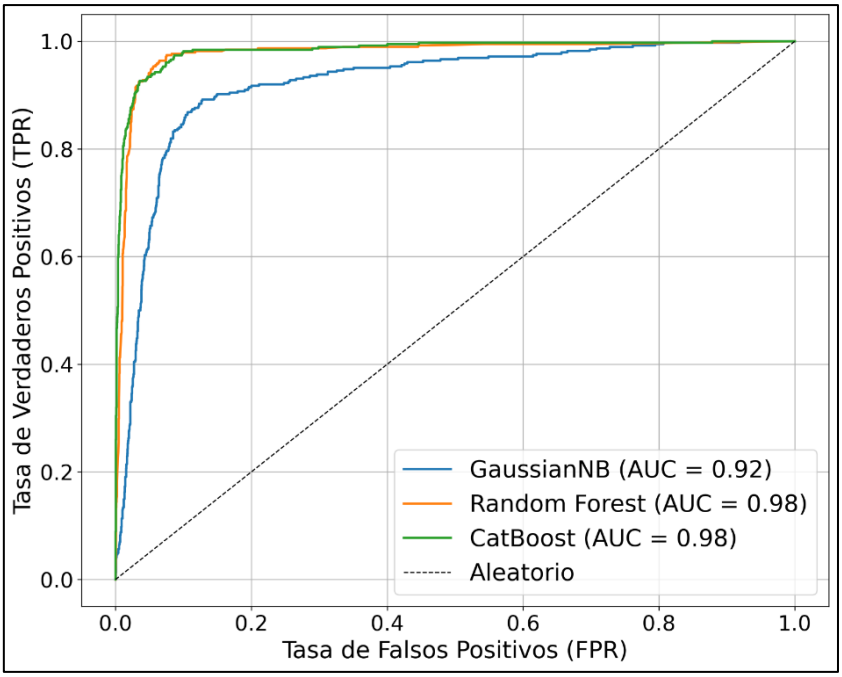

Figura 1. *Curvas ROC de Random Forest, CatBoost y Gaussian Naive Bayes para la detección de deserción estudiantil.*

En la **Figura *2*** se observa que las curvas de CatBoost Classifier y Random Forest se mantienen superiores a 0.8 lo largo de gran parte del rango de recall, esto indica que los modelos son capaces de identificar a una alta proporción de estudiantes que efectivamente abandonan, sin perder precisión en sus predicciones. Además, se destaca que CatBoost Classifier alcanza un promedio de precisión (*Average Precision*, AP) de 0.93, seguido por Random Forest con 0.87, mientras que GaussianNB obtiene apenas 0.62. El alto desempeño de CatBoost en esta métrica sugiere una mayor sensibilidad para detectar patrones relacionados con la deserción estudiantil.

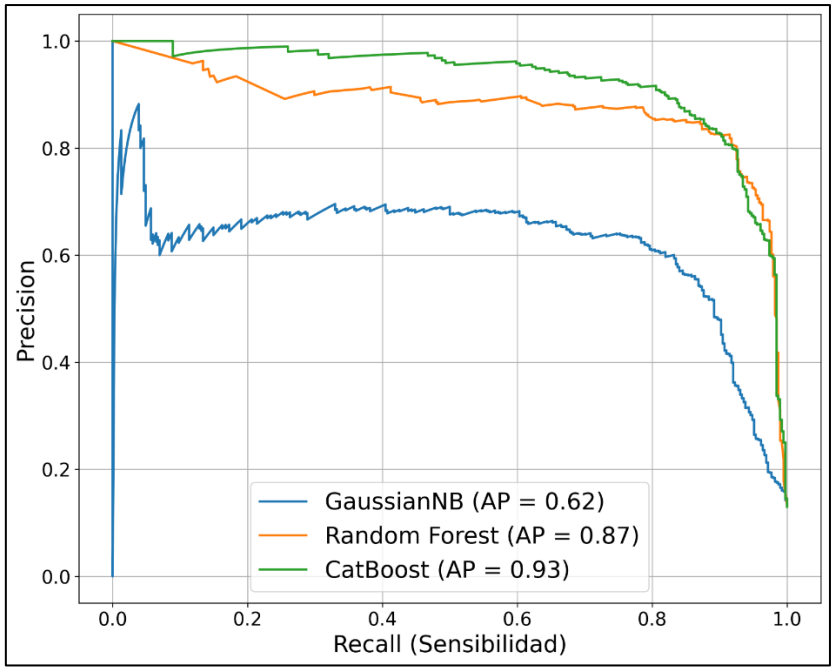

Figura 2. *Curvas Precision-Recall para CatBoost, Random Forest y Gaussian Naive Bayes en la clasificación de deserción estudiantil.*

Por otro lado, la **Figura *3*** presenta las diez variables más relevantes según los modelos Random Forest y CatBoost. En ambos casos, las variables Num_Sus_Temporales, monto_total_girado, GIROS PENDIENTES y EDAD son las de mayor peso predictivo. Estas variables reflejan la importancia de los factores socioeconómicos en la permanencia del estudiante. Además, cabe mencionar que CatBoost Classifier identifica como relevantes variables relacionadas con el puntaje SISBEN y comuna de residencia, esto es debido a que CatBoost puede manejar mejor las variables categóricas, lo que reduce el riesgo de sobreajuste (Vaarma & Li, 2024). Por ello, puede inferirse que CatBoost muestra una mayor sensibilidad ante condiciones de vulnerabilidad de los beneficiarios.

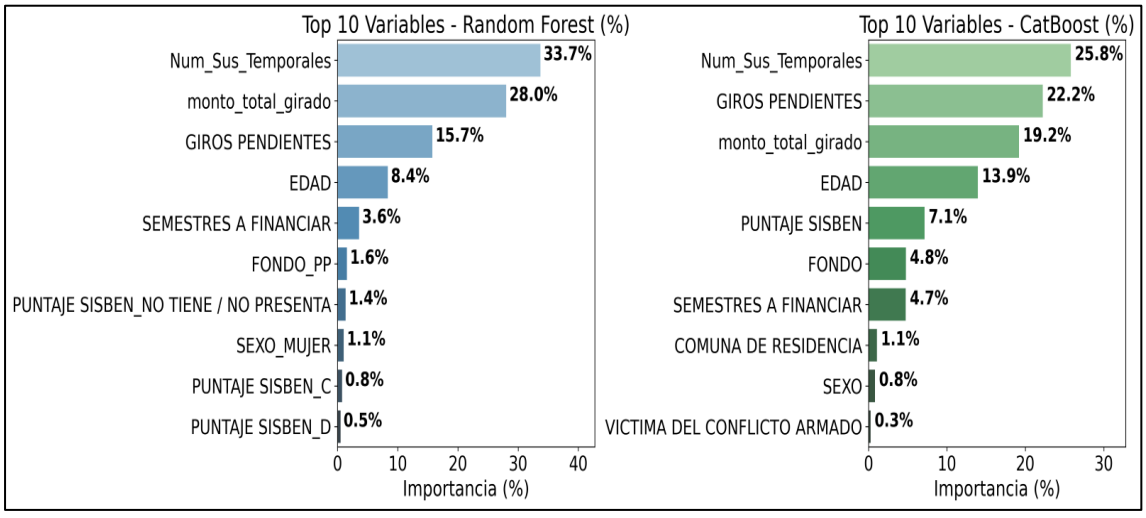

Figura 3. *Variables más relevantes en la predicción de deserción estudiantil según Random Forest y CatBoost.*

Con base en los resultados obtenidos, se identifica que el modelo CatBoost classifier fue el que presentó el mejor desempeño identificando la deserción estudiantil. Su capacidad para manejar variables categóricas de forma nativa y su rendimiento en contextos con clases desbalanceadas le permitieron superar a modelos como Random Forest y Naive Bayes, no solo en términos de precisión global (97.5%), sino en la capacidad de identificar a los desertores (recall del 86%). Estas características hacen de CatBoost la mejor opción entre los modelos evaluados.

3.2. Evaluación crítica

El estudio sobre la deserción postsecundaria en la entidad que ofrece créditos estudiantiles en Medellín presentó limitaciones de alcance, principalmente por la disponibilidad y consistencia de los datos, dado que, apenas desde el año 2019 se virtualizó el proceso de legalización de créditos condonables. Por otro lado, al tratarse de un fenómeno poco frecuente en la entidad, puesto que solo 1906 de los 14641 beneficiarios históricos resultaron ser desertores, los datos estaban fuertemente desbalanceados, lo que hizo que algunos modelos de clasificación mostraran alta sensibilidad al identificar a la población desertora. Para abordar con rigor la complejidad de la deserción estudiantil, es menester incorporar un conjunto de variables más diverso, incluyendo a aliados estratégicos como lo pueden ser las IES a fin de poder analizar esta problemática desde un espectro más amplio.

4. Conclusiones

Mediante el uso de modelos de aprendizaje automático, se logró clasificar en el mejor de los casos con una precisión del 97.5% la deserción estudiantil en los beneficiarios de créditos condonables de pregrado otorgados por una entidad pública de Medellín. Entre los modelos evaluados, CatBoost classifier fue el que obtuvo los mejores resultados con una precisión del 97.5%, gracias a su capacidad para manejar variables categóricas sin necesidad de aplicar métodos adicionales y a su buen rendimiento con clases desbalanceadas, lo que permitió superar a los modelos de Random Forest y Naive Bayes que obtuvieron 99% y 98% respectivamente pero con una tasa de clasificación de verdaderos desertores del 69% y 48% respectivamente, mucho menor al 86% obtenido por el modelo CatBoost classifier .

Las variables más influyentes en los modelos de clasificación fueron el número de suspensiones temporales, giros pendientes, monto total girado, edad y el puntaje SISBEN. Lo que muestra que la combinación de información transaccional, socioeconómica y sociodemográfica permite clasificar con eficacia los casos de deserción estudiantil en esta población.

Los hallazgos de este estudio pueden ser utilizados por la entidad para enfocar sus esfuerzos en la prevención de la deserción en aquellos beneficiarios activos que pueden ser clasificados como desertores por el modelo, con ello fortalecer el área de permanencia de la entidad la cual se encarga de identificar aquellos estudiantes que presentan los riesgos más altos de abandonar sus estudios y persuadirlos de no desistir de sus estudios a través de un acompañamiento constante, además de la articulación con las IES a las que pertenecen estos beneficiarios.

A futuro, se podrían incorporar nuevas fuentes de datos, al igual que actores estratégicos como las IES, que permitan proveer a la entidad de un panorama más amplio a fin de ser más precisos a la hora de identificar a la población potencialmente desertora en la entidad y lograr dilucidar un fenómeno tan complejo y multidimensional como lo es la deserción postsecundaria en Medellín.

Referencias

Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2024). Predicting Student Dropout in Higher Education. *Technology in Society*, 76. doi:https://doi.org/10.1016/j.techsoc.2024.102474

Breiman, L. (2011). Random forests. *45(1)*, 5–32. doi:https://doi.org/10.1023/A:1010933404324

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

Ibrahim, A. A., Ridwan, R. L., Muhammed, M. M., Abdulaziz, R. O., & Saheed, G. A. (2020). Comparison of the CatBoost classifier with other machine learning methods. *International Journal of Advanced Computer Science and Applications*, 11(11). doi:https://doi.org/10.14569/IJACSA.2020.0111190

Kang, K., & Wang, S. (2018). Analyze and predict student dropout from online programs. *ACM International Conference Proceeding*, 6–12.

Martínez, J., & Castillo, D. (2024). Prediction of student dropout using artificial intelligence algorithms. *Procedia Computer Science, 251*, 764–770. doi:https://doi.org/10.1016/j.procs.2024.11.182

Ministerio de Educación Nacional de Colombia. (2023). *Estadísticas de deserción. Sistema de Prevención y Análisis de la Deserción en las Instituciones de Educación Superior (SPADIES)*. doi:https://doi.org/10.14419/ijet.v7i4.44.26862

Pérez, B., Castellanos, C., & Correal, D. (2018). Applying data mining techniques to predict student dropout: A case study. (págs. 1–6). En Proceedings of the 2018 IEEE 1st Colombian Conference on Applications in Computational Intelligence (ColCACI). doi:https://doi.org/10.1109/ColCACI.2018.8484847

Quintero, Y. A. (2022). Diseño de un modelo predictivo para generar alertas tempranas de deserción universitaria en los programas de pregrado presenciales de la Facultad de Ingeniería de la Universidad de Antioquia. *[Tesis de maestría ,Universidad de Antioquia]*. Obtenido de https://bibliotecadigital.udea.edu.co/handle/10495/29368

Sairete, A., Balfagih, Z., Brahimi, T., Mousa, M. E., Lytras, M., & Visvizi, A. (2021). Artificial Intelligence: Towards Digital Transformation of Life, Work, and Education. *Procedia Computer Science, 194*, 1–8. doi:https://doi.org/10.1016/j.procs.2021.11.001

UNESCO. (2020). *Hacia el acceso universal a la educación superior: tendencias internacionales.* París (sede principal de la UNESCO): UNESCO. Obtenido de https://www.iesalc.unesco.org/wp-content/uploads/2020/11/acceso-universal-a-la-ES-ESPANOL.pdf

Vaarma, M., & Li, H. (2024). Predicting student dropouts with machine learning: An empirical study in Finnish higher education. *Technology in Society, 76*(102474). doi:https://doi.org/10.1016/j.techsoc.2024.102474

Vilora, A., Pineda, O. B., & Varela, N. (2019). Bayesian classifier applied to higher education dropout. *Procedia Computer Science*, 573-577. doi:https://doi.org/10.1016/j.procs.2019.11.045