Predicción de retiro de clientes bancarios

Jose Luis Alcocer Cáceres
Juan Carlos Chaverra Bedoya

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

Asesor
Jhon Jair Quiza Montealegre, Magíster (MSc) en Ingeniería

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2023

Cita	Alcocer Cáceres y Chaverra Bedoya [1]
Referencia Estilo IEEE (2020)	[1] J. L. Alcocer Cáceres y J. C. Chaverra Bedoya, “Predicción de retiro de clientes bancarios”, Trabajo de grado especialización, Especialización en Analítica y Ciencia de Datos, Universidad de Antioquia, Medellín, Antioquia, Colombia, 2023.

Especialización en Analítica y Ciencia de Datos Cohorte IV.

Centro de Investigación Ambientales y de Ingeniería (CIA)

Centro de Documentación Ingeniería (CENDOI)

Repositorio Institucional: http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

Dedicatoria

Dedico este trabajo a mis padres, pues sin ellos no hubiera tenido la inspiración y motivación por seguir aprendiendo y superándome cada día y también a cada uno de los profesores del posgrado que de una u otra manera aportaron un granito de arena a la realización de este trabajo. ¡Gracias a todos!

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

SIGLAS, ACRÓNIMOS Y ABREVIATURAS

ML.	Machine Learning.
DL.	Deep Learning.
AI.	Artificial Intelligence.
ROC.	Receiver Operating Characteristic.
AUC.	Area under curve (Área bajo la curva ROC).

RESUMEN

En el mundo empresarial moderno, la fidelización y retención de clientes se han convertido en elementos esenciales y críticos a la hora de definir estrategias y políticas que reduzcan la deserción de clientes hacia otros mercados y/o productos. Perder clientes es más costoso que atraer nuevos. El estudio del comportamiento de los clientes, en particular de su deserción, se ha convertido en una necesidad urgente dentro del ámbito empresarial.

En las empresas financieras, especialmente en los bancos, es un factor crítico entender las deserciones y poder predecir dicho comportamiento. El objetivo principal de este trabajo es encontrar patrones en los datos que permitan identificar y comprender las deserciones, mediante la realización de diferentes iteraciones sobre los datos y utilizando las diferentes técnicas que se abordan en la especialidad de Analítica y Data Science de la Universidad de Antioquia.

El proceso comienza con una primera iteración evaluando los datos a través de un modelo de regresión logística. A partir de ahí, iteraciones posteriores permiten evaluar modelos de aprendizaje automático en busca del modelo óptimo y mejores resultados.

Palabras clave — **Analítica, Ciencia de Datos, Banking, Machine Learning.**

ABSTRACT

In the modern business world, customer loyalty and retention have become essential and critical elements when defining strategies and policies that reduce customer defection to other markets and/or products. Losing customers is more costly than attracting new ones. The study of customer behavior, particularly their defection, has become an urgent necessity within the business sphere.

In financial companies, especially banks, it is a critical factor to understand defections and be able to predict such behavior. The main objective of this work is to find patterns in the data that allow for the identification and understanding of defections, by carrying out different iterations on the data and using the different techniques addressed in the Analytics and Data Science specialization at the University of Antioquia.

The process starts with a first iteration evaluating the data through a logistic regression model. From there, subsequent iterations allow for the evaluation of machine learning models in search of the optimal model and best results.

Keywords — **Analytics, Data Science, Banking, Machine Learning.**

1. INTRODUCCIÓN

En el mundo empresarial moderno la fidelización de los clientes y la retención de estos se han convertido en elementos esenciales y críticos a la hora de definir estrategias y políticas que disminuyan la deserción de clientes hacia otros mercados y/o productos. Perder clientes es más costoso que atraer nuevos clientes [1]. El estudio sobre el comportamiento del cliente, particularmente su deserción, se ha convertido en una necesidad imperante dentro del ámbito empresarial.

En las compañías financieras y especialmente en los bancos es un factor crítico entender las deserciones y poder predecir dicho comportamiento. El objetivo principal del presente trabajo es hallar patrones en los datos que permitan identificar y entender las deserciones, realizando diferentes iteraciones sobres los datos, usando las diferentes técnicas que se abordan en la especialización de Analítica y Ciencia de datos de la Universidad de Antioquia.

Se comienza con un análisis exploratorio de datos y después una primera iteración evaluando los datos a través de un modelo de regresión logística, a partir de allí las iteraciones siguientes permiten evaluar modelos de machine learning en la búsqueda del modelo óptimo y los mejores resultados.

2. PLANTEAMIENTO DEL PROBLEMA

Se aborda la necesidad de las empresas bancarias de retener a sus clientes existentes, ya que adquirir nuevos clientes puede ser muy costoso. Por ende, la predicción del retiro de clientes bancarios se ha convertido en una prioridad para las empresas financieras.

El objetivo es desarrollar un modelo preciso y confiable que pueda ayudar a las empresas bancarias a predecir el retiro de los clientes y tomar acciones a partir de dichos resultados. Para lograr esto, es necesario identificar los factores que influyen en la decisión de retirarse y utilizar técnicas de aprendizaje automático para predecir comportamientos futuros. Por lo tanto, la meta final es identificar las variables más importantes que explican porque un cliente podría darse de baja.

3. JUSTIFICACIÓN

En primer lugar, la retención de clientes es una estrategia fundamental para las empresas bancarias, ya que adquirir nuevos es costoso y puede ser difícil en un mercado altamente competitivo. Por lo tanto, predecir el retiro de los clientes es esencial para tomar medidas preventivas y retener los existentes [1].

¿Pero, esa estrategia debería centrarse únicamente en la retención de clientes existentes? ¿O se podrá abordar un enfoque alternativo para atraer nuevamente a aquellos clientes que ya decidieron retirarse?

Quizás desde un punto de vista empresarial y de minimización de costos, la estrategia a implementar se enfocaría en la retención de clientes existentes siguiendo la premisa que dice adquirir nuevos clientes es más costoso.

En segundo lugar, la utilización de técnicas de aprendizaje automático para predecir el retiro de usuarios se ha vuelto cada vez más común en el sector bancario, pero aún existen desafíos y oportunidades para mejorar la precisión de los modelos de predicción. Es precisamente esta necesidad, la que lleva a la selección del problema planteado, como la oportunidad que existe para analizar un tema conocido desde diferentes puntos de vista con el objetivo de aplicar las mejores técnicas de modelamiento predictivo de datos y algoritmos de machine learning que nos den la información suficiente para la toma de decisiones y elaboración de propuestas y estrategias para la retención de clientes en el sector bancario.

El aporte que este trabajo tendrá a la ciencia es la presentación de un modelo de predicción preciso y confiable para la retención de usuarios en el sector bancario, que puede ser utilizado por las empresas para tomar medidas preventivas y mejorar su capacidad de retener a los usuarios existentes. Además, este trabajo puede servir como una guía para futuras investigaciones en el área de predicción de retiro de usuarios en diferentes sectores.

4. OBJETIVOS

4.1. OBJETIVO GENERAL

Desarrollar un modelo de machine learning para predecir el retiro de clientes con una alta confiabilidad, usando cómo métrica la curva ROC y el área bajo la misma AUC con un valor esperado igual o mayor a 0.8 para poner en producción.

4.2. OBJETIVOS ESPECÍFICOS

- Analizar el comportamiento y relación que tienen las variables independientes con la variable objetivo.
- Implementar diferentes técnicas o algoritmos de aprendizaje supervisado.
- Comparar métricas de diferentes modelos y escoger el más adecuado de acuerdo con el contexto de negocio.
- Interpretar resultados y métricas asociadas a la implementación final.

5. MARCO TEÓRICO

Cuando se piensa en Machine Learning se suele pensar en algo “nuevo” o que ha sido un avance reciente en la ciencia, y la realidad es otra. En los años 60, Frank Rosenblatt, un psicólogo estadounidense de la Universidad Cornell sentó las bases para el campo del aprendizaje profundo quién se basó en sistema nervioso humano para construir una máquina que fuera capaz de reconocer letras del alfabeto, el cual se llamó “perceptrón” [2] y a partir de ahí empezó a tener más relevancia el tema en la ciencia, hasta llegar a lo que percibimos hoy como inteligencia artificial.

Cabe destacar, que el ML, DL y la AI juegan un papel fundamental en la sociedad y el avance de la misma, ejemplo de esto es que tiene aplicaciones en la medicina [3], medio ambiente y contaminación [4], finanzas [5] [6], economía [7], logística [8], entre otras áreas y disciplinas.

Al tener el ML y la AI un campo de aplicación muy amplio, permite implementar soluciones basadas en datos para problemas complejos en las organizaciones, y es acá dónde se acepta el reto de enfrentarse a un problema cotidiano en las compañías prestadoras de servicios como lo es el manejo y retención de los clientes, volviéndose esta la temática a tratar durante el desarrollo de este artículo.

Como resultado del avance del ML a través del tiempo, se tienen muchas soluciones y algoritmos diseñados para resolver problemas en múltiples áreas, es por esto que no debe buscarse solamente la solución que mejor se adapte, sino que también debe buscarse una solución que tenga un buen desempeño, el cual debe ser cuantificado. Algunas métricas para problemas de clasificación son como la precisión (precision), la exhaustividad (recall), el puntaje F1 (f1 score) y la exactitud (accuracy) [9], las cuales suelen ser muy útiles a la hora de escoger un algoritmo de clasificación. De igual manera existen métricas más generales que evalúan el grado de separabilidad entre clases como la curva ROC y el área bajo la misma AUC [10].

6. METODOLOGÍA

La metodología empleada en este estudio se centra en el desarrollo de un modelo de aprendizaje automático capaz de predecir el retiro de clientes bancarios. Para ello, se llevarán a cabo una serie de pasos que permitirán explorar y analizar los datos, preprocesarlos y modelarlos adecuadamente para obtener un buen desempeño predictivo.

Se realizará un análisis exploratorio de datos para analizar patrones y encontrar relaciones entre las variables, y a su vez un preprocesamiento de los datos antes de modelar. Se aplicarán técnicas de normalización, estandarización y codificación de variables categóricas para garantizar que los datos estén en el formato adecuado para el modelado.

Para la modelación se definirá como punto de partida una regresión logística. Esto permitirá obtener una línea base para la comparación con otros modelos que se desarrollen posteriormente.

Se implementarán diferentes modelos y configuraciones de hiperparámetros y se compararán los resultados de los modelos para seleccionar el mejor modelo con el mejor score.

Toda la metodología mencionada se realizará con el lenguaje de programación Python [11].

7. RESULTADOS

7.1 DATOS

Los datos están compuestos por un archivo en formato CSV: *Churn_Modelling.csv*, el cual contiene 14 columnas como se muestra en la tabla 1 con 10.000 observaciones, el peso total del archivo es de 669 Kb.

TABLA 1: TIPOS DE DATOS

Columna	Descripción	Tipo de dato
RowNumber	Número de fila (observación)	int64
Customerld	identificación del cliente	int64
Surname	Apellido	object
CreditScore	Puntaje crediticio	int64
Geography	Lugar de vivienda	object
Gender	Género	object
Age	Edad	int64
Tenure	Tiempo de tenencia	int64
Balance	Balance	float64
NumOfProducts	Número de productos	int64
HasCrCard	Tiene Tarjeta de Crédito	int64
IsActiveMember	Activo	int64
EstimatedSalary	Salario Estimado	float64
Exited	Abandono	int64

7.2. ANÁLISIS EXPLORATORIO

Se comienza llevando a cabo un análisis exploratorio de datos. Esto permite obtener información detallada sobre las características de los datos, detectar posibles patrones y relaciones entre variables, y descubrir cualquier irregularidad que pueda afectar el desempeño del modelo.

Se evidencia que no existen datos faltantes o nulos y se revisan los valores únicos en las variables categóricas para tener certeza de que no haya inconsistencia o errores de escritura que alteren las categorías. Se implementa una exploración visual y estadística de los datos, se calculan las estadísticas descriptivas y las correlaciones entre las variables.

7.2.1 ANÁLISIS DE VARIABLES CATEGÓRICAS

Se observa que el dataset contiene más créditos otorgados en Francia, hay mayoría de personas con tarjetas de crédito, y también que la variable objetivo presenta un desbalanceo, lo cual puede ser un problema al momento de modelar como se observa en la figura A.

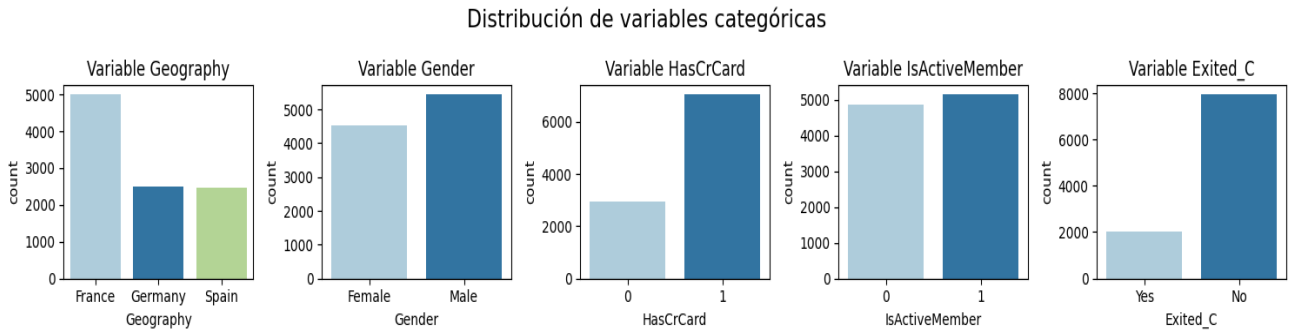

Fig. A Variables categóricas

7.2.2. ANÁLISIS DE VARIABLES NUMÉRICAS

Se realiza estadística descriptiva sobre las variables numéricas y se puede detallar en la figura B que la variable Tenure y NumOfProducts se pueden tratar como variables categóricas, entendiendo que son discretas y provienen de procesos de conteo [12].

	CreditScore	Age	Tenure	Balance	NumOfProducts	EstimatedSalary
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000
mean	650.528800	38.921800	5.012800	76485.889288	1.530200	100090.239881
std	96.653299	10.487806	2.892174	62397.405202	0.581654	57510.492818
min	350.000000	18.000000	0.000000	0.000000	1.000000	11.580000
25%	584.000000	32.000000	3.000000	0.000000	1.000000	51002.110000
50%	652.000000	37.000000	5.000000	97198.540000	1.000000	100193.915000
75%	718.000000	44.000000	7.000000	127644.240000	2.000000	149388.247500
max	850.000000	92.000000	10.000000	250898.090000	4.000000	199992.480000

Fig. B Variables numéricas.

Se implementan gráficos de dispersión con todas las variables numéricas distinguidas por la clase “Exited” en la figura C y en las diagonales se grafican los KDE (Kernel Density Estimate) de cada variable vs la variable respuesta con el fin de observar discriminación de la variable objetivo a través de las variables numéricas.

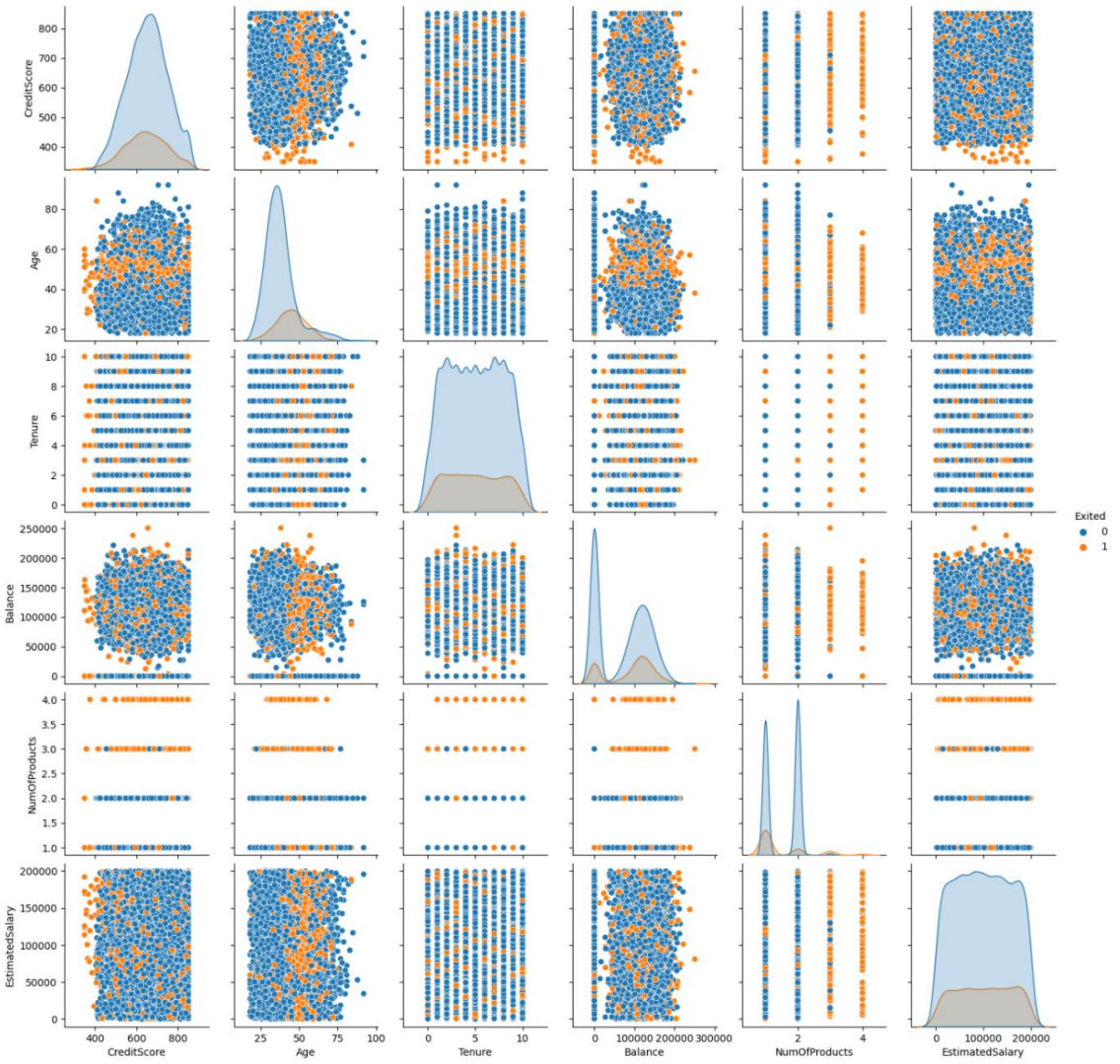

Fig. C Gráficos de dispersión y KDE para variables numéricas.

De la figura anterior se puede decir que la variable “Age” tiene cierto impacto con la variable objetivo, pues parece que la media de edad de quienes no se retiran es menor a la de los que sí lo hacen, y esto puede verse en el diagrama de cajas de la figura D.

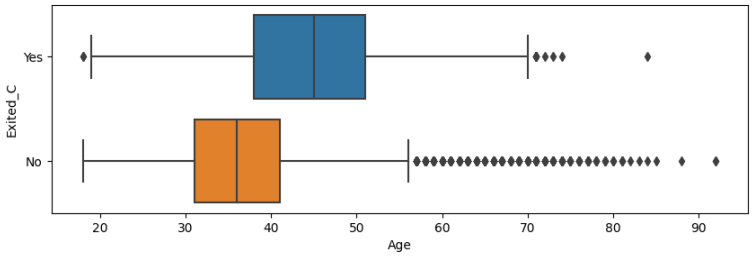

Fig. D Diagrama de cajas para la variable “Age” vs “Exited”.

Por último, se examina las correlaciones entre las variables numéricas para determinar si hay variables que estén altamente correlacionadas y que puedan ser eliminadas para reducir la complejidad del modelo, pero en este caso no existen correlaciones fuertes como se aprecia en la figura E.

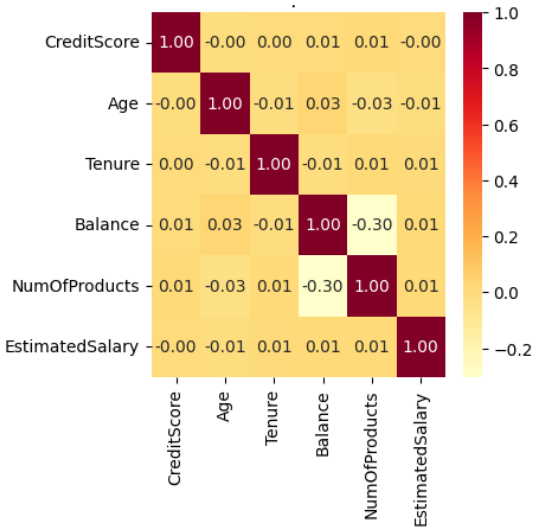

Fig. E Matriz de correlaciones.

7.3 PIPELINE

El pipeline o flujo de trabajo que se propone consta de iteraciones que serán realizadas con el fin de obtener el modelo con el mejor AUC para ser puesto en producción, tal como se ve en la figura F.

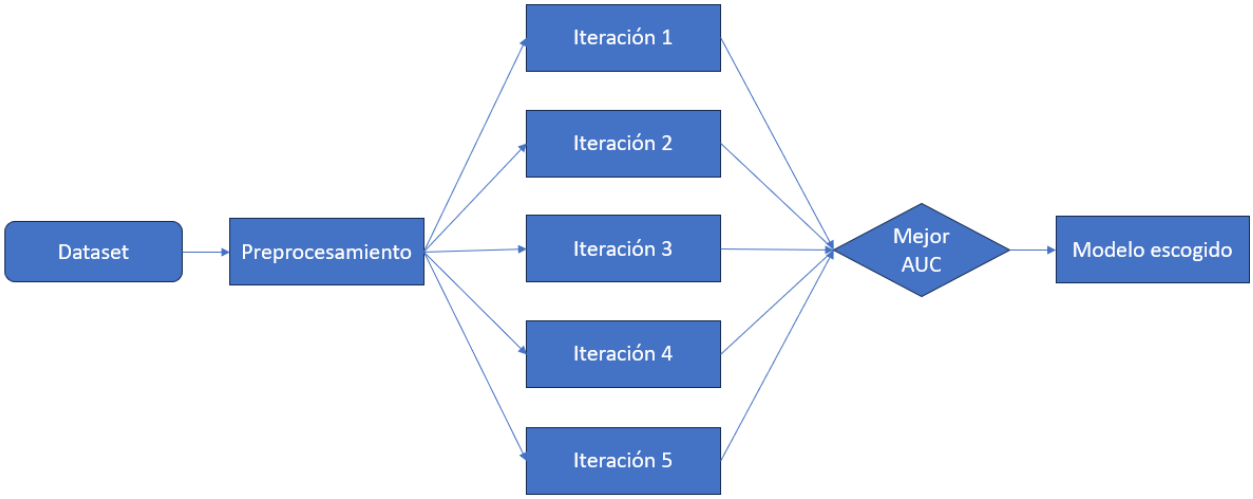

Fig. F Pipeline del proyecto.

7.4 MODELAMIENTO

El modelamiento consta de un preprocesamiento y 5 iteraciones realizadas con el fin de escoger la mejor solución basada en el área bajo la curva ROC. Dentro de esas iteraciones se encuentra el baseline. Los modelos de ML usados pertenecen a la librería de Python llamada *scikit-learn*, la cual tiene diversas aplicaciones para modelamiento, estadística y machine learning [13].

7.5 PREPROCESAMIENTO

Como se observó en la sección de análisis exploratorio, se tienen variables numéricas continuas con diferentes distribuciones y variables categóricas con distintas categorías, por esta razón se escalaron las variables numéricas y se codificaron las variables categóricas, para tener un conjunto de datos listo para un adecuado modelamiento.

Para las variables numéricas se utilizó *MinMaxScaler* y para las categóricas se utilizó OneHotEncoder, ambas de la librería *scikit-learn* [13]. Además, variables como "CustomerId", "Surname", "RowNumber" no fueron tenidas en cuenta ya que no agregan valor a la solución del problema.

Por último, se separan los datos en 2: datos de entrenamiento y prueba. Esto con el fin de poder entrenar con una cantidad de registros y poder evaluar el desempeño con registros que los modelos no hayan visto.

7.6 ITERACIONES

Para cada iteración se usaron los datos de entrenamiento para entrenar los modelos y se calculó el AUC con los datos de prueba para validar desempeño y capacidad de generalización sobre datos no vistos antes, además se utiliza un random state igual a "1234" para garantizar reproducibilidad. Los modelos de machine learning de clasificación que fueron usados son:

- Logistic Regression: Es un algoritmo que utiliza una función logística para estimar la probabilidad de pertenencia a una clase. Ajusta coeficientes mediante la maximización de la función de verosimilitud y clasifica las observaciones según un umbral. Es ampliamente utilizado para problemas de clasificación binaria y multiclase [14].

- Support Vector Machine (SVM): Se basa en la idea de encontrar un hiperplano óptimo que separe las clases de manera óptima en un espacio de mayor dimensión. El algoritmo busca maximizar la distancia entre el hiperplano y los puntos de datos más cercanos, llamados vectores de soporte. SVM utiliza funciones de kernel para mapear los datos a un espacio de características de mayor dimensión y resolver problemas no lineales [15].
- Decision Tree Classifier: Es un algoritmo que crea un árbol de decisión donde cada nodo representa una característica y cada rama una posible decisión. Se divide el conjunto de datos en subconjuntos más puros según las características. Al llegar a las hojas, se asigna una etiqueta de clase a cada instancia. El árbol se utiliza para hacer predicciones clasificando nuevas instancias basándose en las decisiones tomadas durante la construcción del árbol [16].
- Random Forest Classifier: Combina múltiples árboles de decisión. Cada árbol se entrena con una muestra aleatoria del conjunto de datos y utiliza una combinación de características para realizar predicciones. Luego, se promedian las predicciones de todos los árboles para obtener una predicción final. Esto ayuda a reducir el sobreajuste y mejorar la precisión. Además, Random Forest utiliza técnicas como el muestreo bootstrap y la selección aleatoria de características para aumentar la variabilidad y la diversidad entre los árboles del bosque [17].
- KNeighborsClassifier: Está basado en el concepto de vecinos más cercanos. Se asigna una etiqueta de clase a una instancia desconocida basándose en las etiquetas de sus vecinos más cercanos en el espacio de características. El número de vecinos se especifica como un parámetro, y se utiliza una medida de distancia para determinar la cercanía entre las instancias. KNeighborsClassifier se basa en la idea de que las instancias similares tienden a pertenecer a la misma clase y proporciona una forma sencilla pero eficaz de clasificar nuevos datos basándose en su vecindario más cercano [18].

7.6.1: ITERACIÓN 1: BASELINE CON REGRESIÓN LOGÍSTICA

Como línea base se implementó una regresión logística, siendo este debido a su interpretabilidad y eficacia. Como resultado de esta iteración se obtuvo un AUC de 0.75.

7.6.2 ITERACIÓN 2: MÚLTIPLES MODELOS Y AFINAMIENTO DE HIPERPARÁMETROS

Se utilizaron varios modelos y se aplicó búsqueda exhaustiva de los hiperparámetros a cada uno para encontrar los parámetros que mejor AUC generaran, obteniendo los resultados mostrados en la tabla 2.

TABLA 2: RESULTADOS DE LA ITERACIÓN 2.

Modelo	AUC	Mejores parámetros
SVC	0.82	{'C': 5, 'kernel': 'poly'}
DecisionTreeClassifier	0.82	{'max_depth': 10, 'min_samples_split': 100}
RandomForestClassifier	0.86	{'max_depth': 10, 'min_samples_split': 60, 'n_estimators': 100}
KNeighborsClassifier	0.77	{'algorithm': 'ball_tree', 'n_neighbors': 15, 'p': 1}

7.6.3 ITERACIÓN 3: MÚLTIPLES MODELOS, AFINAMIENTO DE HIPERPARÁMETROS Y SELECCIÓN DE CARACTERÍSTICAS

Se implementaron los mismos modelos de la iteración 2, pero esta vez se realiza selección de características a través de las clases SelectKBest, f_classif, chi2 de scikit-learn, las cuales sirven para escoger las variables numéricas y categóricas que tengan mayor relación con la variable objetivo, y así tener un modelo más parsimonioso. Los resultados de esta iteración se ven en la tabla 3.

TABLA 3: RESULTADOS DE LA ITERACIÓN 3.

Modelo	AUC	Mejores parámetros
SVC	0.82	{'C': 5, 'kernel': 'poly'}
DecisionTreeClassifier	0.83	{'max_depth': 10, 'min_samples_split': 100}
RandomForestClassifier	0.86	{'max_depth': 10, 'min_samples_split': 60, 'n_estimators': 100}
KNeighborsClassifier	0.81	{'algorithm': 'brute', 'n_neighbors': 15, 'p': 1}

7.6.4 ITERACIÓN 4: MÚLTIPLES MODELOS, AFINAMIENTO DE HIPERPARÁMETROS Y AJUSTE DE PESOS

Nuevamente se usaron los modelos de las iteraciones anteriores a excepción de KNeighborsClassifier, ya que en esta iteración se ajustan los pesos de las clases a predecir en cada modelo, y KNeighborsClassifier no cuenta con este parámetro. Se realiza esta configuración con el fin de recompensar el desbalanceo de las clases. Los resultados están en la tabla 4.

TABLA 4: RESULTADOS DE LA ITERACIÓN 4.

Modelo	AUC	Mejores parámetros
SVC	0.84	{'C': 5, 'kernel': 'poly'}
DecisionTreeClassifier	0.83	{'max_depth': 10, 'min_samples_split': 100}
RandomForestClassifier	0.86	{'max_depth': 20, 'min_samples_split': 60, 'n_estimators': 150}

7.6.5: ITERACIÓN 5: MÚLTIPLES MODELOS, AFINAMIENTO DE HIPERPARÁMETROS Y BALANCEO DE CLASES

En esta iteración se realizó un balanceo de las clases, ya que como se observa en la figura A, hay más casos con cero de la clase “Exited”, de tal manera que se logró obtener una proporción muy parecida de ambas clases, teniendo 2389 observaciones de la clase 0 y 2037 observaciones de la clase 1. Los resultados se aprecian en la tabla 5.

TABLA 5: RESULTADOS DE LA ITERACIÓN 5.

Modelo	AUC	Mejores parámetros
SVC	0.86	{'C': 5, 'kernel': 'poly'}
DecisionTreeClassifier	0.84	{'max_depth': 10, 'min_samples_split': 100}
RandomForestClassifier	0.87	{'max_depth': 15, 'min_samples_split': 60, 'n_estimators': 100}
KNeighborsClassifier	0.83	{'algorithm': 'ball_tree', 'n_neighbors': 10, 'p': 1}

7.7 MEJOR MODELO

El modelo que mejor desempeño mostró durante las iteraciones fue Random Forest Classifier, el cual obtuvo un AUC de 0.87. Los parámetros establecidos para lograr el mejor desempeño del modelo fueron un max_depth (profundidad máxima) de 10, min_samples_split (división mínima de muestras) de 100, n_estimators (número de árboles) de 100.

7.8 MÉTRICAS ADICIONALES E IMPORTANCIA DE LAS VARIABLES

Se realizaron otras métricas de clasificación para evaluar el rendimiento del modelo escogido, para tener un mayor nivel de detalle y entendimiento de su capacidad predictiva.

7.8.1 PRECISION, RECALL, F1-SCORE

Para el precision, recall y f1-score se obtuvieron los resultados de la tabla 6.

TABLA 6: PRECISION, RECALL Y F1-SCORE.

Clase	Precision	Recall	f1-score
0	0.81	0.86	0.83
1	0.82	0.76	0.79

7.8.2 MATRIZ DE CONFUSIÓN

La matriz de confusión muestra detalladamente cuántas observaciones clasificó bien y mal para cada clase, observándose la cantidad de falsos positivos, falsos negativos, verdaderos positivos y verdaderos negativos, tal como se evidencia en la figura G.

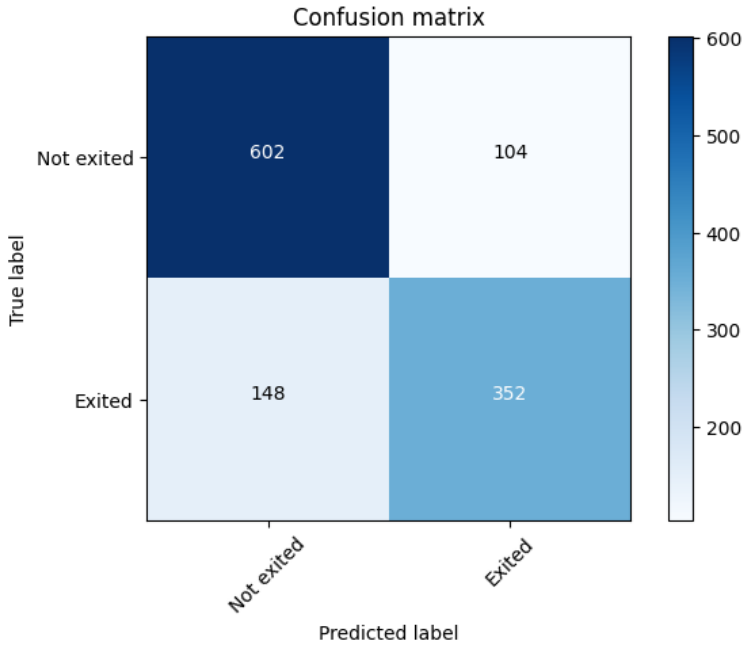

Fig. G: Matriz de confusión.

7.8.3 CURVA ROC

El AUC fue la métrica de clasificación para escoger el mejor modelo, la cual es el área bajo la curva ROC. En la figura H se presenta dicha curva.

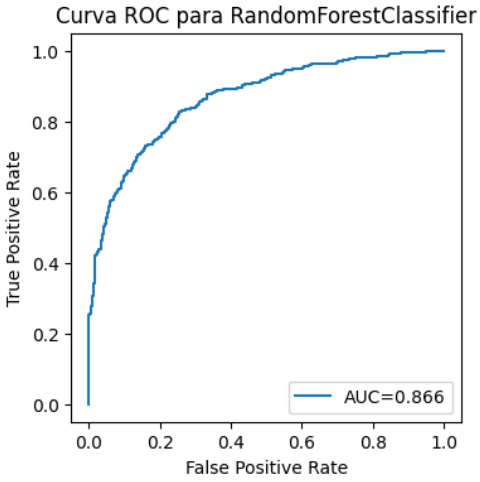

Fig. H: Curva ROC.

7.8.4 IMPORTANCIA DE LAS VARIABLES

La importancia de las variables muestra qué tanto peso o influencia tiene cada una de las variables al momento de realizar una predicción, y para el algoritmo Random Forest Classifier se observa en la figura I.

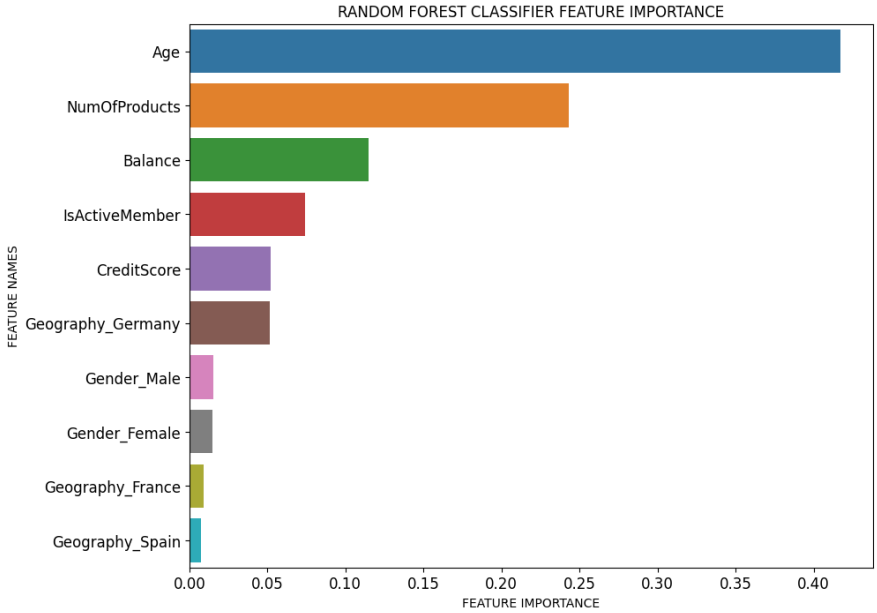

Fig. I Importancia de las variables en el mejor modelo.

8. DISCUSIÓN

A través de la implementación experimental en cada iteración se logró ajustar parámetros y cambiar definiciones que fueron permitiendo poco a poco obtener un mejor resultado y converger a una solución, teniendo una línea base que permitió tener puntos de comparación para escoger la implementación óptima entre todas las opciones.

En la ejecución de cada iteración se implementaron diferentes configuraciones de los modelos, escogiendo los mejores parámetros que optimizaron la métrica AUC y así poder tener varias opciones como solución. En la línea base se obtuvo un AUC de 0.75, lo cual es aceptable, así que de entrada había que encontrar una solución que superara ese umbral. En la iteración 2 se consiguió un mejor AUC respecto a la línea base, el modelo Support Vector Machine y Decision Tree obtienen el mismo valor, Random Forest Classifier el mayor y KNeighbors Classifier el menor como se aprecia en la tabla 2. En la iteración 3 Random Forest Classifier sigue con el mejor AUC y KNeighbors Classifier con el menor a pesar de que aumentó un poco, y Decision Tree Classifier supera a Support Vector Machine como se ve en la tabla 3. Posteriormente, durante las iteraciones 4 y 5, Random Forest Classifier sigue siendo el modelo con mejor AUC, y en ambas iteraciones Support Vector Machine supera a Decision Tree como se observa en las tablas 4 y 5.

Finalmente, en esa última iteración todos los modelos lograron obtener un AUC mayor a 0.8, que era uno de los objetivos de la monografía. Sin embargo, se opta por escoger Random Forest Classifier por tener el mayor AUC. Se observa en la tabla 6 los valores de recall para cada clase, y se interpreta de la siguiente manera: el modelo es capaz de reconocer el 86% de los casos totales de personas que no abandonaron, y el 76% del total de casos que sí lo hicieron.

Las variables más importantes o con mayor peso en el modelo escogido se aprecian en la figura I, teniendo un poco más del 75% de peso las variables edad (40%), número de productos (25%), balance de la cuenta (alrededor de un 12%). La variable edad es la que tiene mayor importancia y en el análisis de variables numéricas, en la figura D, se mostró el impacto que tiene la edad sobre la variable objetivo.

9. CONCLUSIONES

Es interesante la manera en la que se aborda un problema de la vida real a través de la modelación cuantitativa, y el hecho de poder llegar a respuestas o soluciones a problemas complejos a través de disciplinas como la ciencia de datos, demuestra la importancia que esta tiene en múltiples disciplinas.

Se evidenció cómo obtener un resultado aceptable de una solución a un problema de clasificación a través de algoritmos de ML. Esto se llevó a cabo a través de un planteamiento claro del problema y de un flujo de trabajo objetivo.

Se pudo observar un flujo de trabajo estándar en problemas de clasificación, que claramente puede variar dependiendo del problema y los datos disponibles y alcance que se desea obtener con la solución. Se debe dejar clara la importancia de hacer un adecuado análisis exploratorio de datos, limpieza y preprocesamiento de estos, para obtener modelos consistentes.

Cabe destacar también la importancia que tiene el escalamiento de los datos para lograr tiempos de convergencia menores en los algoritmos. Además, la codificación de las variables categóricas permite que los modelos puedan aprender de estas y de su relación con la variable objetivo.

Se concluye que cada uno de los algoritmos de clasificación pueden producir resultados significativamente diferentes en términos de precisión, tiempo de entrenamiento y requisitos computacionales. Por lo tanto, es fundamental evaluar y comparar cuidadosamente múltiples algoritmos para seleccionar el más apropiado para el conjunto de datos y el objetivo específico del proyecto.

10. RECOMENDACIONES

De acuerdo con la implementación realizada, se recomienda usar métodos de búsqueda exhaustiva de hiperparámetros en los algoritmos, pues se evidenció que ayudan a mejorar el rendimiento de los modelos a través de la optimización de las métricas elegidas.

Es una muy buena práctica el hecho de tener un baseline como punto de partida a la hora de resolver problemas de clasificación, puesto que existen muchas soluciones y algoritmos que pueden ajustarse al problema, pero no todos tienen el mismo rendimiento en las mismas métricas. Además, al tener un baseline, permite explorar varias alternativas, comparar entre modelos, y no simplemente implementar un solo modelo y sesgarse frente a los resultados de este.

Es importante mencionar el papel relevante que juega el área bajo la curva ROC (AUC) en los modelos de clasificación, puesto que ofrece una evaluación más robusta y menos sensible al desbalanceo de clases presente en los datos, permitiendo comparar rendimiento entre modelos fácilmente. Por esta razón se recomienda ser tenido en cuenta para este tipo de problemas.

Por último, se recomienda balancear el conjunto de datos cuando se presenta desbalanceo entre clases, ya que se demostró que esto aumenta el rendimiento de los modelos, permitiendo que los algoritmos no tengan más inclinación por una clase que otra y a su vez prevenir el sobreajuste (overfitting).

11. REFERENCIAS

[1] S. V.P., K. Lakshmi, M. Naved, S. K., y V. Podile, «ROLE OF MACHINE LEARNING AND THEIR EFFECT ON BUSINESS MANAGEMENT IN THE WORLD TODAY», vol. 12, pp. 369-374, sep. 2021.

[2] A. L. Fradkov, «Early History of Machine Learning», *IFAC-Pap.*, vol. 53, n.º 2, pp. 1385-1390, ene. 2020, doi: 10.1016/j.ifacol.2020.12.1888.

[3] J. Komuro, D. Kusumoto, H. Hashimoto, y S. Yuasa, «Machine learning in cardiology: Clinical application and basic research», *J. Cardiol.*, may 2023, doi: 10.1016/j.jjcc.2023.04.020.

[4] Y. Li, Z. Sha, A. Tang, K. Goulding, y X. Liu, «The application of machine learning to air pollution research: A bibliometric analysis», *Ecotoxicol. Environ. Saf.*, vol. 257, p. 114911, jun. 2023, doi: 10.1016/j.ecoenv.2023.114911.

[5] Y.-S. Ren, C.-Q. Ma, X.-L. Kong, K. Baltas, y Q. Zureigat, «Past, present, and future of the application of machine learning in cryptocurrency research», *Res. Int. Bus. Finance*, vol. 63, p. 101799, dic. 2022, doi: 10.1016/j.ribaf.2022.101799.

[6] N. Nazareth y Y. V. Ramana Reddy, «Financial applications of machine learning: A literature review», *Expert Syst. Appl.*, vol. 219, p. 119640, jun. 2023, doi: 10.1016/j.eswa.2023.119640.

[7] H. Hala, C. Anass, y B. Youssef, «Machine Learning For the Future Integration of the Circular Economy in Waste Transportation and Treatment Supply Chain», *IFAC-Pap.*, vol. 55, n.º 10, pp. 49-54, ene. 2022, doi: 10.1016/j.ifacol.2022.09.366.

[8] M. D. Capua, A. Ciaramella, y A. De Prisco, «Machine Learning and Computer Vision for the automation of processes in advanced logistics: the Integrated Logistic Platform (ILP) 4.0», *Procedia Comput. Sci.*, vol. 217, pp. 326-338, ene. 2023, doi: 10.1016/j.procs.2022.12.228.

[9] J. Dessain, «Machine learning models predicting returns: Why most popular performance metrics are misleading and proposal for an efficient metric», *Expert Syst. Appl.*, vol. 199, p. 116970, ago. 2022, doi: 10.1016/j.eswa.2022.116970.

[10] A. P. Bradley, «The use of the area under the ROC curve in the evaluation of machine learning algorithms», *Pattern Recognit.*, vol. 30, n.º 7, pp. 1145-1159, jul. 1997, doi: 10.1016/S0031-3203(96)00142-2.

[11] «Welcome to Python.org», *Python.org*, 16 de junio de 2023. https://www.python.org/ (accedido 18 de junio de 2023).

[12] E. Plan, «Modeling and Simulation of Count Data», *CPT Pharmacomet. Syst. Pharmacol.*, vol. 3, n.º 8, p. 129, 2014, doi: 10.1038/psp.2014.27.

[13] «scikit-learn: machine learning in Python — scikit-learn 1.2.2 documentation». https://scikit-learn.org/stable/ (accedido 13 de junio de 2023).

[14] C.-Y. J. Peng, K. L. Lee, y G. M. Ingersoll, «An Introduction to Logistic Regression Analysis and Reporting», *J. Educ. Res.*, vol. 96, n.º 1, pp. 3-14, 2002.

[15] Y. Zhang, «Support Vector Machine Classification Algorithm and Its Application», en *Information Computing and Applications*, C. Liu, L. Wang, y A. Yang, Eds., en Communications in Computer and Information Science. Berlin, Heidelberg: Springer, 2012, pp. 179-186. doi: 10.1007/978-3-642-34041-3_27.

[16] L. Rokach y O. Maimon, «Decision Trees», en *Data Mining and Knowledge Discovery Handbook*, O. Maimon y L. Rokach, Eds., Boston, MA: Springer US, 2005, pp. 165-192. doi: 10.1007/0-387-25465-X_9.

[17] L. Breiman, «Random Forests», *Mach. Learn.*, vol. 45, n.º 1, pp. 5-32, oct. 2001, doi: 10.1023/A:1010933404324.

[18] Z. Zhang, «Introduction to machine learning: k-nearest neighbors», *Ann. Transl. Med.*, vol. 4, n.º 11, p. 218, jun. 2016, doi: 10.21037/atm.2016.03.37.