HERRAMIENTA PARA DETECTAR CLIENTES POTENCIALMENTE FRAUDULENTOS DE BANCOLOMBIA

Santiago Alexis Patiño Munera
Johan Alexis Berrio Arenas

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Directora
Maria Constanza Pabon Burbano

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, JUNIO DE 2025

FICHA RESUMEN
TRABAJO DE GRADO DE MAESTRÍA

TÍTULO**: "HERRAMIENTA PARA DETECTAR CLIENTES POTENCIALMENTE FRAUDULENTOS DE BANCOLOMBIA"**

1. ÁREA DE TRABAJO: Sector Financiero
2. TIPO DE PROYECTO: Aplicado
3. ESTUDIANTES: Santiago Alexis Patiño Munera, Johan Alexis Berrio Arenas
4. CORREO ELECTRÓNICO: sapatino@javerianacali.edu.co, johanberrio@javerianacali.edu.co
5. DIRECCIÓN Y TELEFONO: Av 47c 59-88 Medellín - 3122106847
6. DIRECTORA: Maria Constanza Pabon Burbano
7. VINCULACIÓN DE LA DIRECTORA:
8. CORREO ELECTRÓNICO DE LA DIRECTORA: mcpabon@javerianacali.edu.co
9. CO-DIRECTOR (Si aplica): NA
10. GRUPO O EMPRESA QUE LO AVALA: Bancolombia
11. OTROS GRUPOS O EMPRESAS: NA
12. PALABRAS CLAVE: Fraude externo, monitoreo, clientes, aprendizaje automático, modelo, datos.
13. FECHA DE INICIO: Julio 2024
14. DURACIÓN ESTIMADA (En meses): 12
15. RESUMEN:

En el ámbito bancario, la detección y prevención de fraudes externos es crucial debido a la sofisticación de los métodos empleados por defraudadores. Bancolombia enfrenta el riesgo de fraudes cometidos por clientes, quienes con acceso a servicios y productos que el banco ofrece, pueden realizar actividades ilícitas que impactan económicamente y dañan la reputación de la institución. Los sistemas actuales de monitoreo alertan sobre clientes sospechosos, pero su incapacidad para contextualizar adecuadamente cada cliente resulta en una alta tasa de falsos positivos. El objetivo de este proyecto es desarrollar un modelo de aprendizaje automático para detectar clientes fraudulentos de Bancolombia, integrando datos financieros, transaccionales y demográficos específicos. Con el objetivo de optimizar la asignación de recursos en la investigación de fraudes reales y fortalecer la seguridad financiera de la entidad, se espera obtener los siguientes resultados: una base de datos integrada y equilibrada, un modelo eficiente para la

detección de clientes fraudulentos y un informe detallado que evalúe el desempeño del modelo implementado. La implementación exitosa mitigará los riesgos operativos del fraude externo y promoverá la aplicación de la ciencia de datos para fortalecer la seguridad financiera y la confianza pública en Bancolombia. Además, este proyecto podría servir como referencia para otras entidades, mejorando la eficiencia operativa y reduciendo costos asociados con la gestión de alertas de fraude.

TABLA DE CONTENIDO

FICHA RESUMEN 2

INTRODUCCIÓN 7

1. DEFINICIÓN DEL PROBLEMA 8

1.1. PLANTEAMIENTO DEL PROBLEMA 8

1.2. FORMULACIÓN DEL PROBLEMA 8

2. OBJETIVOS DEL PROYECTO 10

2.1 OBJETIVO GENERAL 10

2.2 OBJETIVOS ESPECÍFICOS 10

2.3 RESULTADOS ESPERADOS 10

3. ALCANCE 11

4. JUSTIFICACIÓN 11

5. MARCO DE REFERENCIA 12

5.1. MARCO TEÓRICO 12

5.1.1 GESTIÓN DEL RIESGO OPERATIVO EN EL CONTEXTO FINANCIERO 12

5.2. ANTECEDENTES 26

6. METODOLOGÍA 28

7. ENTENDIMIENTO DEL NEGOCIO 30

8. ENTENDIMIENTO DE LOS DATOS 35

9. PREPARACIÓN DE LOS DATOS 44

10. BALANCEO DE LOS DATOS 45

11. MODELADO 46

12. EVALUACIÓN 54

13. CONCLUSIÓN 59

14. TRABAJO FUTURO 60

15. REFERENCIAS BIBLIOGRÁFICAS 61

LISTA DE FIGURAS

Figura 1. Análisis de frecuencia de las variables categóricas 39
Figura 2. Boxplot de la variable "Edad" 40
Figura 3. Boxplot de "Edad" por "Tipo de cliente" 41
Figura 4. Matriz de asociación asimétrica entre variables categóricas 43
Figura 5.Porcentaje de varianza explicada por número de componentes 47
Figura 6. Métricas para el modelo KNN con distintos hiperparámetros 49
Figura 7.Métricas para el modelo Random Forest con distintos hiperparámetros.....50
Figura 8. Métricas para el modelo XGBoost con distintos hiperparámetros 51
Figura 9. Métricas para el modelo Decision Tree con distintos hiperparámetros 52
Figura 10. Métricas resultantes del proceso de optimización de hiperparámetros.....57

LISTA DE TABLAS

Tabla 1. Ejemplo de codificación con mask_hash (Fuente: Elaboración propia) 15
Tabla 2. Listado, descripción, posibles valores y tipo de las columnas resultantes 36
Tabla 3. Ejemplo del enmascaramiento realizado a los documentos 38
Tabla 4. Asociaciones relevantes, matriz de Theil (U) 43
Tabla 5. Mejores hiperparámetros para cada modelo 53
Tabla 6. Resultados del proceso de optimización de hiperparámetros para XGBoost.....55

INTRODUCCIÓN

En el ámbito bancario, la detección y prevención de fraudes constituye un desafío crucial debido a la constante evolución de métodos sofisticados empleados por defraudadores. De particular preocupación son los fraudes perpetrados por clientes, quienes, con acceso a productos y servicios que el banco ofrece, pueden perpetrar actividades ilícitas que pueden involucrar a la organización financiera. Estas acciones no solo representan pérdidas económicas significativas, sino que también impactan adversamente la reputación y la confianza pública en la institución.

Bancolombia, consciente de estas amenazas, ha implementado sistemas automatizados de monitoreo para alertar sobre clientes sospechosos. Sin embargo, la eficacia de estos sistemas se ve limitada por su incapacidad para contextualizar adecuadamente cada cliente, lo que conduce a una alta tasa de falsos positivos, superiores al 90%. Esta deficiencia radica en la implementación basado en reglas duras y falta de consideración del contexto individual de cada cliente, como historial transaccional, hábitos de pago y datos demográficos, al momento de evaluar la legitimidad de los clientes.

En el presente proyecto de grado se desarrolló un modelo de aprendizaje automático para detectar clientes fraudulentos de Bancolombia. El objetivo es integrar datos financieros, transaccionales y demográficos específicos de cada cliente en un sistema de monitoreo, optimizando así la asignación de recursos para la investigación de fraudes reales. Con el apoyo tecnológico y estratégico de Bancolombia, este proyecto busca fortalecer la seguridad financiera de la entidad y mantener su integridad frente a los constantes desafíos del fraude externo.

Este trabajo exploró técnicas avanzadas de aprendizaje automático como XGBoost, Random Forest y k-Nearest Neighbors (KNN), además de estrategias de balanceo de datos como Synthetic Minority Over-sampling Technique (SMOTE) y Generative Adversarial Networks (GANs). Estas técnicas se aplican para enfrentar el desbalance en la distribución de clases, característico por la naturaleza poco frecuente pero crucial del fraude externo de clientes. En resumen, este proyecto no solo buscó mitigar los riesgos operativos asociados con el fraude externo, sino también avanzar en la aplicación práctica de la ciencia de datos para fortalecer la seguridad financiera y la confianza pública en las instituciones bancarias como Bancolombia.

1. DEFINICIÓN DEL PROBLEMA

1.1. PLANTEAMIENTO DEL PROBLEMA

En el sector bancario, los fraudes evolucionan rápidamente, y los defraudadores utilizan métodos y herramientas cada vez más sofisticados para mover dinero de forma ilícita o robar directamente a las entidades financieras [1]. Un agente potencial de fraude se encuentra dentro de los consumidores de las entidades: sus clientes directos. Con acceso a productos y servicios y, en algunos casos conocimiento de sus vulnerabilidades y tipologías de fraude, ciertos clientes podrían verse tentados a realizar actividades financieras ilícitas. Además de los daños económicos directos, la reputación de las organizaciones se ve gravemente afectada cuando sus clientes se involucran en procesos ilícitos externos como lavado de dinero, narcotráfico y extorsión. Las actividades criminales de cara a un banco no solo ponen en riesgo la integridad financiera de la entidad, sino que también erosiona la confianza de los clientes y del público en general, afectando negativamente la imagen y la credibilidad de la institución [2].

Actualmente, en Bancolombia existen sistemas de monitoreos automáticos diseñados para alertar sobre movimientos atípicos en las cuentas de los clientes con el objetivo de identificar posibles defraudadores. El problema central radica en la ineficacia del sistema actual de detección de clientes fraudulentos, que produce una alta tasa de falsos positivos superior al 90%. Este sistema está basado en reglas duras que no consideran el contexto individual de cada cliente, tratándolos de manera generalizada sin tener en cuenta el historial transaccional, indicadores de riesgo, la información demográfica y el momento de vida de cada cliente.

La necesidad de un enfoque más eficiente y preciso requiere la exploración de un modelo que evalúe los perfiles en función del contexto individual de cada cliente. La implementación de una herramienta de este tipo podría reducir los falsos positivos, dirigiendo los recursos de investigación hacia alertas de riesgo real. Esto optimizaría el uso de recursos, fortalecería la seguridad financiera y mitigaría el impacto en la reputación de Bancolombia.

1.2. FORMULACIÓN DEL PROBLEMA

¿Cómo desarrollar un sistema de detección de clientes potencialmente fraudulentos de Bancolombia?

Para responder la pregunta de investigación es necesario contestar los siguientes interrogantes: ¿Cómo integrar la información financiera, transaccional y demográfica de cada cliente en el sistema de monitoreo? ¿Cómo ajustar la base para manejar eficazmente datos desbalanceados y con pocas etiquetas positivas? ¿Qué técnica de la Ciencia de Datos puede emplearse para

determinar el perfil fraudulento o no fraudulento de un cliente? ¿Qué métricas pueden emplearse para evaluar los resultados?

2. OBJETIVOS DEL PROYECTO

2.1 OBJETIVO GENERAL

Implementar un modelo de aprendizaje automático que permita la detección de clientes potencialmente fraudulentos de Bancolombia.

2.2 OBJETIVOS ESPECÍFICOS

- Crear una base de datos anonimizada que recopile el contexto personal de cada cliente, abarcando indicadores de riesgo, su historial transaccional y datos demográficos.
- Aplicar técnicas de balanceo de datos, asegurando una representación adecuada de las distintas categorías y variables, para evitar sesgos y mejorar la robustez del análisis.
- Entrenar un modelo de aprendizaje automático con los datos recopilados de cada cliente.
- Evaluar el desempeño del modelo midiendo su efectividad en la detección de clientes fraudulentos.

2.3 RESULTADOS ESPERADOS

- Base de datos anonimizada e integrada con las características relevantes de los clientes.
- Base de datos balanceada en términos de la proporción entre clientes fraudulentos y no fraudulentos.
- Modelo de aprendizaje automático para la detección de clientes fraudulentos.
- Informe del desempeño del modelo.

3. ALCANCE

El proyecto se desarrolló con datos de clientes de Bancolombia de 2011 en adelante, garantizando la anonimización de la identificación de los clientes involucrados para evitar faltas éticas en la manipulación y exposición de su información real. Se exploró el uso de características financieras, como productos obtenidos y comportamiento crediticio, y demográficas, como edad y región. Dado que la tasa de fraude externo es menor al 27% del total de alertas generadas, se cuenta con un dataset desbalanceado en cuanto a clientes fraudulentos. Para abordar esto, se emplearon técnicas de balanceo de datos con el objetivo de lograr una mejor proporción de cantidad de clientes fraudulentos y no fraudulentos en la base de datos final. Debido a la ausencia de un modelo de aprendizaje automático de referencia en este proceso, se entrenaron cuatro algoritmos en Python para su posterior evaluación, considerando tanto los datos recopilados de investigaciones previas como el análisis y opinión de un experto. Finalmente, es importante destacar que implementar estos modelos en producción implica un proceso organizacional que va más allá del alcance de este proyecto.

El proyecto tuvo el aval de la empresa Bancolombia, la cual aportó la infraestructura tecnológica, acompañamiento directivo y acceso a las bases de datos necesarias para la realización del proyecto. Sin embargo, por políticas de la Organización no se permitió extraer, compartir o manipular los datos fuera de la infraestructura asignada por el Banco. El directivo encargado de acreditar la realización del proyecto en la Organización hizo parte del área de Fraude Externo y tenía total conocimiento del proceso original.

4. JUSTIFICACIÓN

El sector bancario enfrenta constantemente el desafío de detectar y prevenir fraudes externos en constante evolución, los cuales no solo generan pérdidas económicas significativas sino también dañan la reputación de las instituciones financieras. La implementación de este proyecto optimiza el uso de recursos al dirigir eficazmente las investigaciones hacia potenciales clientes fraudulentos. Esto permite a Bancolombia mejorar la protección de sus activos financieros y mantener la confianza de sus clientes y transparencia en sus prácticas financieras. Además, teniendo en cuenta que el no realizar este tipo de monitoreos y generación de alertas puede ocasionar sanciones a Bancolombia por parte de los entes reguladores, este proyecto garantiza el cumplimiento de dicha normatividad y, por ende, evita la ejecución de dichas amonestaciones financieras.

Resultados positivos en este tipo de predicciones posicionaría a Bancolombia como una entidad innovadora en la implementación de tecnologías avanzadas para la detección de fraudes. Además, podría servir como referencia y ser aplicado en otras entidades, contribuyendo de manera más amplia a la prevención de fraudes en distintos sectores. Por otro lado, una implementación de

este tipo mejoraría la eficiencia operativa y reducirá los costos asociados con la gestión de alertas de esta índole.

5. MARCO DE REFERENCIA

5.1. MARCO TEÓRICO

Este apartado expone los conceptos clave y el marco normativo relacionados con la detección de fraude externo en entidades financieras, y describe las técnicas y herramientas identificadas para abordar este problema. En el contexto de este estudio, se pretende monitorear si los clientes están involucrados en actividades ilícitas a través de su información financiera y demográfica.

5.1.1 GESTIÓN DEL RIESGO OPERATIVO EN EL CONTEXTO FINANCIERO

Todas las entidades bajo la supervisión y control de la Superintendencia Financiera de Colombia (SFC) están obligadas a implementar un Sistema de Administración de Riesgo Operativo (SARO). Este sistema tiene como objetivo identificar, medir, controlar y monitorear de manera eficaz los riesgos operativos [3].

Riesgo Operativo (RO): Se define como la posibilidad de incurrir en pérdidas debido a deficiencias, fallas o inadecuaciones en los recursos humanos, procesos, tecnología, infraestructura o por la ocurrencia de acontecimientos externos [4].

En este trabajo, se busca mitigar el siguiente riesgo operativo:

Fraude Externo: Actividad ilícita cometida por personas ajenas a una organización, con el fin de obtener beneficios económicos [5].

A continuación, se presentan las definiciones de conceptos fundamentales para el entendimiento del contexto normativo y operativo relacionado al Fraude Externo.

SARLAFT: Es el Sistema de Administración del Riesgo de Lavado de Activos y de la Financiación del Terrorismo, una herramienta implementada principalmente en el sector financiero y otras entidades reguladas en Colombia para identificar, evaluar, prevenir, controlar y mitigar riesgos

relacionados con el lavado de activos y la financiación del terrorismo. El SARLAFT está regulado por la Superintendencia Financiera de Colombia (SFC), bajo normativas como la Circular Básica Jurídica (Circular Externa 029 de 2014 y actualizaciones). Estas regulaciones exigen a las entidades aplicar un enfoque basado en riesgo para prevenir actividades ilícitas [6].

ROS: En el contexto de la prevención del lavado de activos y la financiación del terrorismo, un ROS es un Reporte de Operación Sospechosa. Este término se refiere al informe que las entidades financieras y otros sujetos obligados deben elaborar y presentar ante las autoridades competentes cuando detectan actividades o transacciones inusuales que podrían estar relacionadas con lavado de activos o financiación del terrorismo [7].

Criterios para Presentar un ROS:

- Transacciones que no tienen justificación económica o comercial.
- Operaciones repetitivas o fragmentadas para evitar controles.
- Comportamientos inusuales del cliente (evasión de preguntas, resistencia a proporcionar documentación, entre otros).

Listas de control: Las listas de control son herramientas utilizadas en el contexto de la prevención de lavado de activos y financiación del terrorismo (LA/FT), así como en otros ámbitos de cumplimiento normativo, para identificar personas, entidades, o jurisdicciones asociadas con actividades ilícitas o consideradas de alto riesgo [8]. Las listas de control incluyen nombres de personas, empresas, organizaciones o países vinculados a:

- Actividades criminales (narcotráfico, terrorismo, corrupción, etc.).
- Sanciones económicas o comerciales impuestas por organismos internacionales o gobiernos.
- Riesgo alto en materia de lavado de activos y financiación del terrorismo.

PEP (Persona Políticamente Expuesta): Una Persona Expuesta Políticamente es alguien que ocupa o ha ocupado cargos públicos relevantes, o tiene relaciones cercanas con personas en esos cargos. Por ejemplo:

- Presidentes, ministros, diputados, senadores
- Altos cargos en el poder judicial o militares
- Ejecutivos de empresas estatales
- Familiares o personas con relaciones cercanas a estas personas

El concepto se usa principalmente en el ámbito financiero y legal para identificar personas que, por su posición, podrían estar en mayor riesgo de estar involucradas en actos de corrupción, lavado de dinero o financiamiento ilícito. Por eso, las instituciones financieras y otras

organizaciones suelen aplicar controles más estrictos para evitar riesgos asociados al manejo de dinero de estas personas [9].

Persona Natural

Una persona natural es un ser humano con capacidad para adquirir derechos y contraer obligaciones. Es decir, es cualquier individuo en su condición de sujeto de derecho [10].

Características:

- Tiene derechos y deberes civiles.
- Puede celebrar contratos, adquirir bienes, ser responsable legalmente, etc.
- Su capacidad puede estar limitada por la ley.

Persona Jurídica

Una persona jurídica es una entidad creada por una o varias personas naturales, reconocida por la ley, que tiene derechos y obligaciones propias, distintas de las de sus integrantes. Es decir, es una "persona" que no es humana, sino una organización o empresa [10].

Características:

- Puede ser una empresa, asociación, fundación, sociedad, etc.
- Tiene patrimonio propio.
- Puede celebrar contratos, ser demandada o demandar, poseer bienes y asumir responsabilidades legales.
- Su existencia depende del cumplimiento de ciertos requisitos legales.

5.1.2 BASES DE DATOS RELACIONALES

Debido que la información de los clientes se encontraba almacenada en tablas pertenecientes a bases de datos relacionales, fue necesario estructurar consultas en lenguaje SQL que permitieron integrar, consolidar y anonimizar adecuadamente los datos requeridos para el análisis.

Bases de datos relacionales: Una base de datos relacional organiza datos en filas y columnas dentro de tablas, donde existen datos interrelacionados. Los datos se estructuran en múltiples tablas, conectadas mediante claves primarias o externas. Estos identificadores únicos muestran las relaciones entre tablas y se representan mediante el Modelo Relacional [11].

SQL: El Lenguaje de Consultas Estructuradas (SQL), es un lenguaje de programación diseñado para interactuar con datos almacenados en sistemas de gestión de bases de datos relacionales. Su uso

es ampliamente extendido en diversas aplicaciones, que van desde sistemas empresariales como MySQL, PostgreSQL, Oracle Database y Microsoft SQL Server, hasta plataformas orientadas al análisis de datos y Big Data. Entre sus principales ventajas se destacan la estandarización de los sistemas, la flexibilidad en las operaciones de manipulación y consulta de datos, así como su capacidad de interoperabilidad con otros lenguajes de programación y herramientas utilizadas en ciencia de datos [12].

mask_hash: es una función integrada en SQL Hive que se utiliza para enmascarar datos sensibles. Esta función toma como entrada un valor y devuelve un hash irreversible, lo que significa que no es posible recuperar el valor original. Es especialmente útil para anonimizar datos como identificaciones personales, números de tarjeta de crédito o cualquier información confidencial que deba protegerse.

Tabla 1. Ejemplo de codificación con mask_hash (Fuente: Elaboración propia)

id	nombre	identificacion_real	identificacion_enmascarada
1	Juan	123456789	f4a12bcd89e567aef23c345d98
2	Ana	987654321	d89f7bcd1234abc5e67def2345
3	Luis	567890123	b1234ab567f98dcef4571234ab

5.1.3 ANÁLISIS EXPLORATORIO DE DATOS

Rango intercuartílico: (IQR, por sus siglas en inglés) es una medida estadística que describe la dispersión de un conjunto de datos al enfocarse en la distancia entre el primer cuartil (Q1) y el tercer cuartil (Q3) de la distribución [13].

$$IQR = Q3 - Q1$$

- Q1 (Primer cuartil): Es el valor que delimita el 25% inferior de los datos ordenados.
- Q3 (Tercer cuartil): Es el valor que delimita el 75% inferior de los datos ordenados.

El IQR representa el rango dentro del cual se encuentra el 50% central de los datos. Se usa en la identificación de valores atípicos y estos se identifican cuando se encuentran por fuera del rango:

$$[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$$

Matriz de asociación basada en Theil's U: es una herramienta utilizada para cuantificar la dependencia direccional entre variables categóricas, basada en los principios de la teoría de la información. Esta matriz se construye a partir del coeficiente U de Theil, una métrica que mide cuánto se reduce la incertidumbre (entropía) de una variable aleatoria Y al conocer otra variable X. [14].

- El Coeficiente U de Theil se define como:

$$U(Y|X) = \frac{H(Y) - H(Y|X)}{H(Y)}$$

Este valor está acotado entre 0 y 1:

  - $U(Y|X) = 0 : X$ no proporciona ninguna información sobre Y
  - $U(Y|X) = 1$: El conocimiento X elimina completamente la incertidumbre sobre Y

- La entropía marginal de una variable Y se calcula como:

$$H(Y) = -\sum_{y \in Y} P(y) \log_2 P(y)$$

Donde $P(y)$ es la probabilidad de ocurrencia de cada valor y de la variable Y

- La entropía condicional de Y dado X se define como:

$$H(Y|X) = -\sum_{x \in X} P(x) \sum_{y \in Y} P(y|x) \log_2 P(y|x)$$

Donde:

  - $P(x)$ es la probabilidad marginal de x
  - $P(y|x)$ es la probabilidad condicional de y dado x

- Matriz de asociación Theil's U: se construye evaluando el coeficiente U para todas las combinaciones posibles de variables categóricas V_i y V_j. La entrada [i, j] de la matriz se define como:

$$\text{Matriz}_{i,j} = U(V_j \mid V_i) = \frac{H(V_j) - H(V_j \mid V_i)}{H(V_j)}$$

Es importante notar que esta matriz no es simétrica, es decir, en general:

$$U(V_j|V_i) \neq U(V_i|V_j)$$

La matriz de asociación Theil's U se utiliza ampliamente en el análisis exploratorio de datos, ya que permite:

- Detectar relaciones de dependencia direccional entre variables categóricas.
- Evaluar el poder explicativo de una variable respecto a otra.
- Realizar selección de variables o reducción de dimensionalidad.
- Interpretar estructuras de dependencia en modelos predictivos.

5.1.4 TÉCNICAS ESTANDARIZACIÓN DE DATOS

Dado que los datos de los clientes presentaban distintas tipologías y escalas, fue necesario aplicar métodos de estandarización que facilitaran un modelado adecuado. Para ello, se emplearon las técnicas de One-Hot Encoding y Standard Scaler, que permiten transformar variables categóricas y normalizar valores numéricos, respectivamente.

One-Hot Encoding: Es una técnica de preprocesamiento utilizada para convertir variables categóricas en un formato numérico que pueda ser interpretado por algoritmos de machine learning. Consiste en crear una columna binaria (con valores 0 o 1) para cada categoría posible de la variable, indicando con un 1 la presencia de una categoría en un registro y con 0 su ausencia. Esta codificación evita que el modelo asuma un orden o relación matemática entre las categorías [15].

Standard Scaler: Es una técnica de normalización que transforma los datos para que tengan una media igual a 0 y una desviación estándar igual a 1. Se utiliza comúnmente en algoritmos de machine learning que son sensibles a la escala de las variables, como SVM o regresión logística. La fórmula aplicada es:

$$z = \frac{x - \mu}{\sigma}$$

donde x es el valor original, μ la media de la variable y σ su desviación estándar. Esta transformación asegura que todas las variables contribuyan de forma equitativa al modelo [16].

5.1.5 TÉCNICAS DE BALANCEO DE DATOS

Dado el desbalance presente en el conjunto de datos utilizado en el proyecto, fue necesario implementar técnicas de balanceo con el fin de mitigar posibles sesgos y mejorar el rendimiento de los modelos predictivos. En este contexto, se exploraron dos enfoques ampliamente utilizados en la literatura: SMOTE (Synthetic Minority Over-sampling Technique) y Redes Generativas Antagónicas (GANs)

Balanceo de datos: Son estrategias utilizadas para abordar el desequilibrio en la distribución de clases en un conjunto de datos. Cuando tenemos clases desequilibradas, es decir, una clase minoritaria con muy pocas muestras en comparación con una clase mayoritaria, los algoritmos de aprendizaje automático pueden verse afectados en su capacidad para generalizar correctamente [17].

SMOTE: Synthetic Minority Over-sampling Technique, es una técnica ampliamente utilizada en problemas de clasificación con conjuntos de datos desbalanceados, donde una clase tiene significativamente menos instancias que otra. En lugar de simplemente duplicar las instancias existentes de la clase minoritaria, SMOTE genera nuevos ejemplos sintéticos al interpolar entre las muestras existentes de esa clase. Este método se basa en seleccionar un punto aleatorio entre una instancia minoritaria y uno de sus vecinos más cercanos, lo que permite crear datos que conservan la distribución general de la clase minoritaria y, al mismo tiempo, evitan problemas como el sobreajuste asociado al muestreo repetitivo.

Los pasos básicos de SMOTE son:

1. Identificación de vecinos más cercanos: Para cada instancia de la clase minoritaria, se identifican sus k vecinos más cercanos en el espacio de características.
2. Interpolación: Se selecciona aleatoriamente uno de estos vecinos y se calcula un punto intermedio en el espacio vectorial.
3. Generación de muestras sintéticas: Este punto intermedio se utiliza como una nueva instancia de la clase minoritaria.

SMOTE es especialmente efectivo cuando se combina con técnicas de aprendizaje automático que son sensibles a clases desbalanceadas, mejorando así el rendimiento del modelo en términos de métricas como la sensibilidad, la *precision* y el área bajo la curva ROC (AUC). Además, SMOTE puede ser ajustado mediante hiperparámetros como el número de vecinos (k_neighbors) para controlar el nivel de sobremuestreo [18].

Redes Generativas Antagónicas: GANs, por sus siglas en inglés, son una poderosa técnica para la generación de datos sintéticos, que se puede utilizar para el balanceo de conjuntos de datos desequilibrados en aprendizaje automático. Las GANs consisten en dos redes neuronales principales que compiten entre sí:

- Generador (Generator): Su objetivo es generar datos sintéticos que sean indistinguibles de los datos reales.
- Discriminador (Discriminator): Su objetivo es distinguir entre los datos reales y los datos generados por el generador.

Estas dos redes se entrenan de manera simultánea en un proceso de competencia (juego min-max):

El generador intenta mejorar sus habilidades para producir datos sintéticos realistas.

El discriminador intenta mejorar sus habilidades para identificar correctamente los datos reales de los datos sintéticos [19].

5.1.6 TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

Para detectar patrones sospechosos en información financiera, se exploraron técnicas de aprendizaje automático, una rama de la inteligencia artificial que utiliza datos y algoritmos para imitar el aprendizaje humano [20]. En este proyecto se consideraron varios métodos: SVM, Random Forest, KNN, XGBoost y Decision Tree, aunque finalmente se aplicaron Random Forest, KNN, XGBoost y Decision Tree.

SVM: Las Máquinas de Soporte Vectorial (SVM, por sus siglas en inglés) son una poderosa técnica de aprendizaje supervisado ampliamente utilizada para la clasificación, regresión y detección de anomalías. En el contexto de la identificación de clientes fraudulentos, las SVM son especialmente útiles debido a su capacidad para manejar problemas de alta dimensionalidad y encontrar fronteras de decisión óptimas entre clases [21].

Random Forest: Es un algoritmo de clasificación basado en árboles de decisión y técnicas de ensamble. Su objetivo es mejorar la precisión y estabilidad del modelo al combinar múltiples clasificadores débiles en un clasificador más robusto.

El modelo genera un conjunto de árboles de decisión $h_1(x)$, $h_2(x)$, …, $h_T(x)$, cada uno entrenado sobre un subconjunto aleatorio del conjunto de datos original, mediante muestreo con reemplazo (bootstrap). Además, en cada división de los nodos, se selecciona aleatoriamente un subconjunto de características, lo que introduce diversidad adicional entre los árboles.

La predicción final se realiza mediante votación mayoritaria entre los árboles del bosque. Dada una instancia de entrada x, la clase predicha se define como:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \ldots, h_T(x)\}$$

Donde $h_T(x)$ representa la predicción del árbol número t y T es el número total de árboles.

Cada árbol, en su proceso de entrenamiento, selecciona divisiones que maximizan la pureza de las clases en los nodos. Una de las métricas más utilizadas para esto es el índice de Gini, definido como:

$$Gini(D) = 1 - \sum_{i=1}^{C} p_i^2$$

Donde:

D es el conjunto de datos en el nodo actual, C es el número de clases, p_i es la proporción de instancias pertenecientes a la clase i dentro de D [22].

- Hiperparámetros Esenciales

Algunos de los hiperparámetros más relevantes para ajustar un modelo Random Forest para clasificación son:

1. *n_estimators*: Número de árboles en el bosque
2. *max_depth:* Profundidad máxima de cada árbol
3. *min_samples_split:* Mínimo número de muestras requerido para dividir un nodo
4. *min_samples_leaf:* Número mínimo de muestras requerido para estar en una hoja
5. *max_features:* Número de características consideradas al buscar la mejor división
6. *bootstrap:* Indica si se utiliza muestreo con reemplazo

K-Nearest Neighbors (KNN): Es un algoritmo supervisado de aprendizaje automático utilizado tanto para tareas de clasificación como de regresión. Se basa en el principio de similitud, clasificando o prediciendo valores para nuevas instancias en función de la cercanía a los datos existentes en el espacio de características. KNN es conocido por su simplicidad, robustez y capacidad para adaptarse a problemas no lineales sin requerir un modelo paramétrico explícito [23].

- Fundamentos Matemáticos

  En KNN, la predicción para una instancia desconocida x se realiza identificando los k vecinos más cercanos en el espacio de características. La distancia entre x y cada punto x_i del conjunto de entrenamiento se calcula utilizando métricas como la distancia euclidiana:

$$d(x, x_i) = \sqrt{\sum_{j=1}^{n} \left(x_j - x_{i,j}\right)^2}$$

  Donde n es el número de características.

1. Clasificación: La etiqueta se asigna por mayoría entre los k vecinos más cercanos.

$$\hat{y} = \text{mode}\{y_i : i \in N_k(x)\}$$

  Donde $N_k(x)$ es el conjunto de los k vecinos más cercanos de x.

- Hiperparámetros Esenciales

  KNN permite ajustar varios hiperparámetros clave para mejorar su rendimiento:

  1. *k (n_neighbors):* Define el número de vecinos considerados para clasificar o predecir una instancia. Un valor pequeño puede conducir a un modelo más sensible al ruido, mientras que un valor grande puede diluir la influencia de los vecinos más cercanos.
  2. *Metric (métrica de distancia):* Especifica la métrica utilizada para calcular la cercanía, como la distancia euclidiana, Manhattan o Minkowski.
  3. *Weighting (pesos):* Determina si todos los vecinos tienen la misma influencia (uniformuniformuniform) o si se ponderan de acuerdo con su distancia (distancedistancedistance).

XGBoost: eXtreme Gradient Boosting, es un algoritmo basado en árboles de decisión que utiliza el enfoque de gradient boosting, diseñado específicamente para optimizar el rendimiento en tareas de clasificación y regresión. Este método es ampliamente reconocido por su eficiencia computacional, robustez y capacidad para manejar datos de alta dimensionalidad y complejidad. XGBoost se basa en el principio de ensamble, donde se combinan secuencialmente múltiples árboles de decisión para corregir los errores de predicción de los árboles anteriores y mejorar continuamente la *precision* del modelo [24].

- Fundamentos Matemáticos

  En XGBoost, la predicción final se obtiene como una combinación ponderada de los resultados de los árboles $f(x)$:

$$f(x) = \sum_{k=1}^{k} f_k(x)$$

  Donde $f_k(x)$representa el k-ésimo árbol de decisión, y K es el número total de árboles. Durante el proceso de entrenamiento, el objetivo es minimizar una función de pérdida $L(y, \hat{y})$ penalizada por la complejidad del modelo:

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

  donde $\Omega(f_k)$ incluye términos de regularización que ayudan a prevenir el sobreajuste.

- Hiperparámetros Esenciales

  XGBoost ofrece un amplio control sobre el modelo mediante una variedad de hiperparámetros. Los más relevantes incluyen:

  4. *colsample_bytree:* Este hiperparámetro especifica la proporción de características que se seleccionarán de forma aleatoria para construir cada árbol.
  5. *learning_rate (η):* Controla la contribución de cada árbol al modelo final. Valores más bajos conducen a un aprendizaje más lento, pero tienden a mejorar la generalización.
     - Fórmula de actualización de predicción:

$$F_{t+1}(x) = F_t(x) + \eta f_t(x)$$

     donde $F_t(x)$ es el modelo acumulado hasta la iteración t.

  6. *max_depth:* Define la profundidad máxima de cada árbol de decisión, controlando la complejidad del modelo.
  7. *n_estimators:* Determina el número total de árboles en el ensamble. Un número mayor de árboles puede mejorar la *precision*, pero aumenta el tiempo de entrenamiento.
  8. *subsample:* Representa la fracción de muestras utilizadas para entrenar cada árbol, seleccionadas aleatoriamente del conjunto de datos. Esto ayuda a reducir el riesgo de sobreajuste.

Decision Tree (Árboles de Decisión): Es un modelo de aprendizaje supervisado que divide iterativamente los datos en subconjuntos basados en características, utilizando criterios de decisión como la ganancia de información o el índice Gini. Este enfoque es popular debido a su interpretabilidad y facilidad de uso tanto para tareas de clasificación como de regresión. [25]

- Fundamentos Matemáticos
  1. Criterio de División

  Los Árboles de Decisión evalúan cada división posible en un nodo utilizando medidas como:

  a) Índice Gini (G):

$$G = 1 - \sum_{i=1}^{C} p_i^2$$

donde p_i es la proporción de ejemplos pertenecientes a la clase i en el nodo actual.

b) Ganancia de Información (IG):

$$IG = H(\text{parent}) - \sum_{j=1}^{k} \frac{|D_j|}{|D|} H(D_j)$$

donde H es la entropía y D_j son los subconjuntos generados por la división.

2. Entropía (H):

$$H(D) = -\sum_{i=1}^{C} p_i \log_2(p_i)$$

3. Clasificación:
   El nodo terminal asigna la clase mayoritaria de las muestras que contiene:

$$\hat{y} = \text{mode}\{y_i : i \in \text{leaf}(x)\}$$

- Hiperparámetros Esenciales

1. *max_depth:* Controla la profundidad máxima del árbol, limitando su complejidad para prevenir sobreajuste.
2. *min_samples_split:* Define el número mínimo de muestras necesarias para dividir un nodo.
3. *min_samples_leaf:* Especifica el número mínimo de muestras que debe contener un nodo hoja.
4. *criterion:* Determina la métrica utilizada para evaluar la calidad de una división (e.g., 'gini', 'entropy').
5. *max_features:* Número máximo de características consideradas para dividir un nodo.

5.1.7 TÉCNICAS DE OPTIMIZACIÓN

En el marco del proyecto, se exploraron diversas técnicas de optimización que son fundamentales para mejorar el desempeño de los modelos predictivos en problemas de clasificación. Entre las técnicas empleadas destacan el Análisis de Componentes Principales (PCA), Random Search y Grid Search, las cuales se describen a continuación.

PCA (Análisis de Componentes Principales): Es una técnica de reducción de dimensionalidad que transforma un conjunto de variables originales correlacionadas en un nuevo conjunto de variables

no correlacionadas, denominadas componentes principales. Estas componentes se ordenan de forma que las primeras explican la mayor proporción de la varianza presente en los datos.

En resumen, el proceso que sigue el algoritmo de PCA hace lo siguiente:

a) Estandarización de los datos:
Para garantizar que todas las variables contribuyan de manera equitativa al análisis, se estandarizan los datos. Esto implica convertirlos en una escala común mediante la fórmula:

$$Z = \frac{X - \mu}{\sigma}$$

donde X es la matriz de datos, μ el promedio y σ la desviación estándar de cada variable.

b) Cálculo de los autovalores y autovectores:
Se obtiene la matriz de covarianza C de los datos estandarizados, y posteriormente se realiza la descomposición espectral:

$$Cv_i = \lambda_i v_i$$

donde λ_i son los autovalores y $\boldsymbol{v_i}$ los autovectores.

c) Selección de componentes principales:
Los autovalores se ordenan en orden descendente para identificar las direcciones principales de mayor varianza. Se seleccionan los k autovectores asociados a los k mayores autovalores, donde k representa las dimensiones del nuevo subespacio reducido.

d) Construcción de la matriz de proyección:
Se forma la matriz de proyección W combinando los k autovectores seleccionados como columnas:

$$W = [v_1, v_2, \dots, v_k]$$

e) Transformación de los datos:
Los datos originales estandarizados se proyectan en el nuevo subespacio de características utilizando W:

$$X_{\text{PCA}} = ZW$$

El resultado es una representación de los datos en un espacio de k dimensiones que preserva la mayor parte de la varianza [26].

Random Search: Es una técnica de optimización de un modelo que selecciona de manera aleatoria combinaciones de hiperparámetros dentro de un espacio predefinido. En lugar de evaluar todas

las combinaciones posibles, se enfoca en una muestra aleatoria, lo que permite cubrir de forma eficiente un amplio rango de configuraciones. El proceso de selección puede describirse como una búsqueda estocástica:

$$\theta_i \sim U(\Theta)$$

donde θ_i es un conjunto de hiperparámetros seleccionado aleatoriamente, y Θ representa el espacio total de posibles combinaciones [27].

Grid Search: Es un método exhaustivo de búsqueda en el que se evalúan todas las combinaciones posibles de hiperparámetros dentro de un rango predefinido. Este enfoque garantiza encontrar la configuración óptima, aunque es computacionalmente más costoso. El cálculo matemático es:

$$\prod_{i=1}^{n} |\Theta_i|$$

Si $\Theta_1, \Theta_2, \ldots, \Theta_n$ son los valores discretos definidos para n hiperparámetros, el Grid Search evalúa las combinaciones posibles, donde $|\Theta_i|$ representa el número de valores posibles para el hiperparámetro i [28].

5.1.8 MÉTRICAS DE EVALUACIÓN

Con el objetivo de obtener resultados cuantitativos, se emplearon diversas métricas para la evaluación de los algoritmos de clasificación. Para medir el desempeño de los modelos desarrollados, se utilizaron *Precision*, *Recall*, F1-Score y *Accuracy*.

Precision: Es la proporción de instancias correctamente clasificadas como positivas (verdaderos positivos) respecto al total de instancias clasificadas como positivas (verdaderos positivos y falsos positivos). Es especialmente relevante en problemas donde el costo de los falsos positivos es alto, como en la detección de fraudes [29].

$$Precision = \frac{\text{Verdaderos Positivos (VP)}}{\text{Verdaderos Positivos (VP)} + \text{Falsos Positivos (FP)}}$$

Recall (Sensibilidad o Tasa de Verdaderos Positivos): Mide la capacidad del modelo para identificar correctamente todas las instancias positivas. Es la proporción de verdaderos positivos respecto al total de instancias que realmente son positivas (verdaderos positivos y falsos negativos) [29].

$$Recall = \frac{\text{Verdaderos Positivos (VP)}}{\text{Verdaderos Positivos (VP)} + \text{Falsos Negativos (FN)}}$$

F1-Score: Es la media armónica entre *precision* y *recall*, proporcionando una única métrica para evaluar el balance entre estos dos indicadores. Es útil cuando existe un desbalance en las clases o cuando es importante considerar tanto los falsos positivos como los falsos negativos.

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Un F1-Score cercano a 1 indica un buen desempeño del modelo tanto en términos de *precision* como de *recall* [29].

Accuracy (Exactitud): Mide la proporción de instancias correctamente clasificadas (tanto positivas como negativas) respecto al total de instancias. Aunque es una métrica sencilla de interpretar, puede ser engañosa en conjuntos de datos desbalanceados, ya que un modelo podría tener alta exactitud clasificando todas las instancias como la clase mayoritaria.

$$Accuracy = \frac{\text{Verdaderos Positivos (VP)} + \text{Verdaderos Negativos (VN)}}{\text{Total de Instancias}}$$

En problemas de clasificación desbalanceada, métricas como F1-Score o *Recall* suelen ser más informativas que el *Accuracy* [29].

5.2. ANTECEDENTES

A continuación, se presentan los trabajos relacionados que se identificaron en la búsqueda de referentes y estado del arte de la identificación de clientes fraudulentos mediante herramientas de aprendizaje automático:

- Aplicación de Modelos de Aprendizaje Automático en la Detección de Fraudes en Transacciones Financieras: Este proyecto evalúa el desempeño de técnicas como Random Forest y redes neuronales para detectar transacciones financieras en una base de datos real. Los resultados presentan un excelente comportamiento de identificación para ciertos modelos con rendimientos sobre el 90%, esta investigación presenta una gran utilidad en el desarrollo de este proyecto pues da un punto de partida en la selección de técnicas y un posible paso a paso para evaluar los resultados de estas [30].

- Detección de anomalías en el ámbito del fraude financiero: Este artículo se encarga de investigar y presentar una revisión exhaustiva de las técnicas de detección de anomalías más populares y efectivas aplicadas para detectar fraudes financieros, con un enfoque en resaltar los avances recientes en las áreas de aprendizaje semi-supervisado y no supervisado. Además, presenta los limitantes identificados en la detección mediante aprendizaje supervisado, información muy relevante para este proyecto, puesto que se tiene un avance en la estructuración de este tipo de soluciones. El proyecto concluye que los sistemas de detección, especialmente basados en técnicas recientes de aprendizaje semi-supervisado y no supervisado, han demostrado ser efectivos para identificar anomalías y reducir el impacto económico del fraude [31].

- Credit Card Fraud Detection Using Minority Oversampling and Random Forest Technique: Este trabajo utiliza distintos métodos de aprendizaje automático como arboles de Decisión (DT), k Vecinos Más Cercanos (KNN), Regresión Logística (LR) y Bosque Aleatorio (RF) para detectar fraudes en tarjetas de crédito. El dataset utilizado presenta 284,807 de las cuales 492 fueron encontradas como fraudulentas. Debido al gran desbalanceo de los datos, se aplica la técnica de SMOTE para crear datos sintéticos llegando a 284,315 transacciones legítimas y 284,315 fraudulentas. La conclusión luego de entrenar y evaluar los modelos es que el Random Forest presenta las mejores métricas para este caso [32].

- A Data Mining Based System For Transaction Fraud Detection: Este artículo adopta un sistema semiautomático de detección de fraude en transacciones basado en Random Forest y detección manual en la que un experto revisa los posibles fraudes. El dataset utilizado contiene más de 1 millón de muestras, cada una con más de 400 variables características, tanto financieras como no financieras. Finalmente, la *precision* del modelo alcanzó el 96.8% y la puntuación AUC ROC fue de 92.5% [33].

- Application of GA Feature Selection on Naive Bayes, Random Forest and SVM for Credit Card Fraud Detection: El estudio compara la efectividad de la Máquina de Soporte Vectorial (SVM) y los Sistemas Basados en Reglas (RBS) para detectar fraudes en transacciones financieras. Utiliza un conjunto de datos financieros diverso que incluye múltiples tipos de transacciones, volúmenes y características relevantes, todas etiquetadas como legítimas o fraudulentas. Los resultados muestran que los modelos de aprendizaje automático, como la SVM, alcanzan una puntuación F1 del 90%, una *precision* del 95% y una exactitud del 92%. Sin embargo, este rendimiento superior viene acompañado de una mayor necesidad de recursos computacionales, con un tiempo de procesamiento de 120 milisegundos [34].

6. METODOLOGÍA

Para alcanzar los objetivos propuestos, se utilizó la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), una metodología estándar y ampliamente empleada en la minería de datos, que proporciona un proceso estructurado y bien definido para proyectos de análisis de datos y aprendizaje automático. CRISP-DM se compone de seis fases iterativas principales: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación. Esta metodología permite que el equipo avance a través de las diferentes fases de manera flexible, regresando a etapas anteriores cuando sea necesario (por ejemplo, del modelado a la preparación de datos) y definiendo hitos útiles a lo largo del proyecto para asegurar su éxito [35].

Actividades:

En esta sección, se enumeran las principales actividades realizadas durante el desarrollo del proyecto. Cada actividad representa una etapa clave del proceso y contribuye al logro de los objetivos planteados. A continuación, se presenta una lista de estas actividades, las cuales serán descritas en detalle en las secciones posteriores.

1. Entendimiento del Negocio

   Aunque por motivos de confidencialidad no es posible divulgar ciertos aspectos específicos del negocio, se presenta de manera general cómo opera la organización, su estructura y la forma en que se articula el trabajo interdisciplinario entre equipos. En particular, se describe la composición del equipo encargado de la gestión del riesgo de fraude externo, su rol dentro del proceso y su relación con otras áreas estratégicas.

   Durante esta etapa, se realizó un análisis exhaustivo en colaboración con los expertos internos, especialmente el equipo de Fraude Externo, con el objetivo de comprender el contexto, los objetivos y los principales desafíos asociados al riesgo de fraude externo. Este trabajo conjunto permitió delimitar con claridad el problema a resolver, definir los alcances del proyecto, alinear expectativas con los recursos disponibles y, adicionalmente, recomendar un conjunto inicial de variables clave para el desarrollo del modelo de detección.

2. Entendimiento de los datos

   - Recolección de datos de las fuentes identificadas.
   - Integración de las tablas.
   - Anonimización de la base.
   - Exploración inicial de los datos para entender su estructura y contenido.

- Identificación de problemas de calidad de datos (valores faltantes, duplicados, inconsistencias).
- Análisis exploratorio de datos (EDA) para identificar patrones y relaciones relevantes.

3. Preparación de los datos

    - Integración de diferentes fuentes de datos en un único dataset.
    - Limpieza y preprocesamiento de datos (normalización, transformación, manejo de valores faltantes).

4. Balanceo de datos
    - Aplicación de técnicas de balanceo de datos como SMOTE o GANs para abordar el desbalanceo de clases.

5. Modelado

    - Selección de algoritmos de aprendizaje automático (XGBoost, Random Forest, KNN, Decision Tree).
    - División de los datos en conjuntos de entrenamiento y prueba.
    - Entrenamiento de múltiples modelos utilizando los datos preparados.
    - Ajuste de hiperparámetros para optimizar el rendimiento de los modelos.

6. Evaluación

    - Evaluación de los modelos entrenados utilizando métricas como *Precision*, *Recall*, F1-Score y *Accuracy*.
    - Comparación de los modelos para seleccionar el más efectivo.
    - Preparación de un informe detallado sobre el desempeño del modelo.

7. Despliegue (Trabajo futuro)

    - Implementación del modelo seleccionado en el entorno de producción.
    - Integración del modelo con los sistemas de monitoreo y alerta existentes.

7. ENTENDIMIENTO DEL NEGOCIO

Bancolombia, en su calidad de entidad financiera regulada por la Superintendencia Financiera de Colombia, tiene la obligación de identificar, medir, controlar y monitorear de manera eficaz los riesgos operativos. En este contexto, el presente proyecto tiene como objetivo principal mitigar el riesgo operativo asociado al fraude externo, mediante el desarrollo de un modelo de detección de clientes potencialmente fraudulentos.

Para alcanzar este objetivo, la organización cuenta con una estructura organizacional clara e interdisciplinaria, que permite la colaboración efectiva entre diferentes equipos especializados. En el marco de este proyecto, los principales actores involucrados son:

1. Equipo de Fraude Externo

Este equipo tiene la responsabilidad de centralizar, diseñar, desarrollar y ejecutar los modelos de detección de fraude relacionados con clientes de la organización. Su principal objetivo es identificar patrones y comportamientos asociados a distintas tipologías de fraude, en respuesta a las necesidades planteadas por diferentes áreas de la organización, conocidas internamente como gerencias.

El presente proyecto fue planteado como una evolución del esquema de monitoreo previamente utilizado, el cual se basaba en reglas duras para generar alertas sobre posibles clientes fraudulentos. En su lugar, se propuso el desarrollo de un modelo analítico que permitiera una detección más precisa y robusta. Los objetivos y el alcance del proyecto fueron definidos y alineados desde el inicio, permitiendo así una articulación efectiva entre los frentes académico y laboral.

Debido a la sensibilidad de la información y a las políticas internas de tratamiento de datos, el uso de los datos fue aprobado exclusivamente dentro de la infraestructura tecnológica de la organización. Por lo tanto, no fue posible exportar información a fuentes externas como servidores en la nube, repositorios como Git, o dispositivos locales. Asimismo, toda la información empleada fue debidamente anonimizada para proteger la identidad de los clientes.

Como parte del proceso de desarrollo del modelo, se llevaron a cabo diversas sesiones de trabajo entre los analistas e investigadores del equipo de Fraude Externo y los desarrolladores del modelo. En una primera etapa, se consolidó una base de datos de etiquetas de fraude/no fraude, asignadas por expertos. Es importante destacar que, debido a la sensibilidad del tema, cada etiqueta se trató como un caso independiente, y un mismo cliente pudo haber sido evaluado en distintos momentos, generando múltiples registros para un mismo individuo.

A partir de estas mesas se definió un conjunto de variables relevantes para el modelo, las cuales fueron agrupadas en cuatro grandes categorías:

A. Variables Corporativas

El desarrollo del modelo comenzó con la incorporación de una serie de variables construidas internamente por la organización. Estas variables, conocidas como corporativas, fueron creadas a partir de aprendizajes obtenidos a lo largo de investigaciones previas y del análisis de casos históricos relacionados con fraude.

Es importante aclarar que, por razones de confidencialidad, no es posible explicar en detalle los criterios específicos que se utilizaron para definir ciertos umbrales o clasificaciones dentro de ellas. Estos criterios responden a lineamientos internos y al conocimiento experto del equipo, y su divulgación está restringida por políticas de seguridad de la información.

Las variables corporativas son:

- Recién vinculado: Esta variable identifica si un cliente ha sido vinculado a la organización en los últimos 24 meses. Se considera relevante debido a que, en los primeros meses de relación, la entidad aún no cuenta con suficiente información transaccional para evaluar su comportamiento. Además, investigaciones previas han evidenciado patrones de fraude asociados a clientes con poca antigüedad.
- Persona jurídica recién constituida: Evalúa si una persona jurídica fue legalmente constituida hace seis meses o menos. La reciente creación de estas entidades puede representar un mayor riesgo debido a la falta de historial financiero y de comportamiento.
- Segmento: Es una clasificación comercial realizada por el equipo de conocimiento del cliente. Esta variable es crítica, ya que el comportamiento transaccional varía significativamente entre segmentos. Por ejemplo:
    - Empresas del sector construcción pueden clasificarse como “CONSTRUCTOR”,
    - Grandes corporaciones como “CORPORATIVO”,
    - Pequeñas empresas como “PYME” o “MICROPYME”,
    - Personas naturales como “PERSONAS”, “SOCIAL”, “PLUS” o “PREFERENCIAL”,
    - Trabajadores independientes como “INDEPENDIENTES” o “NEGOCIOS & INDEPENDIENTES”.
- Exportador: Señala si el cliente ha declarado que se dedica a la exportación de bienes y servicios, es decir, a comercializar productos o servicios producidos localmente en mercados extranjeros.
- Importador: Identifica si el cliente se dedica a la importación, es decir, a traer al país productos o servicios producidos en el exterior para su comercialización interna.
- Empleado: Dado que la primera línea de defensa frente al fraude son los propios colaboradores de la organización, es relevante saber si el cliente es un empleado activo de Bancolombia.
- Región: Clasifica al cliente según su ubicación geográfica, agrupando departamentos en grandes regiones. Por ejemplo:
    - La región CARIBE incluye departamentos como Atlántico y Magdalena.

- La región SUR agrupa zonas como Nariño y Cauca.

B. Variables Demográficas

Por otro lado, se incluyen una serie de variables demográficas, las cuales están relacionadas con las características propias de cada cliente. Estas variables ofrecen contexto adicional sobre el tipo de persona o entidad con la que se establece la relación comercial. Las variables demográficas consideradas son:

- Tipo de cliente: Clasifica al cliente como persona natural (PN) o persona jurídica (PJ). Esta distinción es clave, ya que ambos tipos de clientes presentan comportamientos financieros y patrones transaccionales significativamente distintos.
- Edad: Corresponde a la edad del cliente al momento de la evaluación o etiquetado. En el caso de personas jurídicas, pueden presentarse edades inferiores a un año (cuando son empresas recién constituidas) o superiores a los 100 años (en el caso de entidades con larga trayectoria). Para personas naturales, también se han observado casos de menores de edad a quienes sus tutores legales les han abierto cuentas desde el nacimiento, con el fin de recibir depósitos o administrar recursos.
- Tipo de documento: Indica el tipo de documento de identidad con el cual el cliente fue vinculado a la organización. Algunos ejemplos comunes son: “CÉDULA DE CIUDADANÍA”, “NIT”, “PASAPORTE” y “CÉDULA DE EXTRANJERÍA”.
- PEP (Persona Expuesta Políticamente): Señala si el cliente ocupa, o ha ocupado, un cargo público destacado que lo clasifica como persona expuesta políticamente, lo cual implica un nivel de riesgo superior debido a posibles actos de corrupción o abuso de poder.
- Relacionado PEP: Identifica si el cliente mantiene un vínculo (hasta segundo grado de consanguinidad o afinidad) con una persona expuesta políticamente, lo cual también puede representar un riesgo adicional.
- Mayor: Variable binaria que indica si el cliente ha superado la edad de pensión establecida en Colombia al momento de la evaluación.
- Menor: Variable binaria que determina si el cliente no ha alcanzado la mayoría de edad legal en Colombia (18 años), condición que puede implicar restricciones operativas o necesidades especiales de supervisión.

C. Variables transaccionales

Adicionalmente, se consideraron variables transaccionales, las cuales son calculadas individualmente para cada cliente con base en su segmento comercial y su transaccionalidad a través de cualquier canal (como transferencias electrónicas, corresponsales bancarios, sucursales físicas, entre otros). Esta clasificación es crucial, ya que los distintos segmentos presentan comportamientos financieros significativamente diferentes. Por ejemplo, no se espera el mismo volumen de transacciones en un cliente del segmento "MICROPYME" que en uno del segmento "CORPORATIVO".

Las variables transaccionales utilizadas en el modelo son:

- Supera perfil transaccional: Indica si el cliente se encuentra dentro del percentil 99 en términos del monto total de dinero enviado o recibido en comparación con los clientes de su mismo segmento comercial.
- Supera enviados: Señala si el cliente se ubica en el percentil 99 del total de dinero enviado en relación con el comportamiento promedio de su segmento. Esta variable busca identificar comportamientos atípicos que puedan representar un riesgo.
- Supera recibidos: Refleja si el cliente supera el percentil 99 del monto total de dinero recibido, también en comparación con los demás clientes del mismo segmento comercial.

Estas variables permiten detectar desviaciones significativas frente al comportamiento esperado, lo que contribuye a identificar posibles anomalías transaccionales que podrían estar relacionadas con actividades fraudulentas.

Adicionalmente, se incorporaron variables desarrolladas por el equipo de Gestión del Riesgo, con el fin de complementar el modelo desde una perspectiva especializada en identificación y evaluación de riesgos asociados al cliente.

2. Equipo de Gestión del Riesgo

El equipo de Gestión del Riesgo es responsable de centralizar, definir, calificar y supervisar las características que puedan estar asociadas a posibles riesgos de los clientes. Las variables que este equipo construye se basan en metodologías internas, cuya documentación detallada es estrictamente confidencial y no se divulga fuera del equipo.

Para este proyecto, se incluyeron diversas variables provenientes de este equipo, con el objetivo de enriquecer el análisis del riesgo desde una perspectiva integral. A continuación, se describen las variables incluidas:

- Cliente Sensible: Clasificación del cliente según su nivel de sensibilidad para la organización, en función de factores demográficos, laborales y reputacionales. Las categorías posibles son: *Alta*, *Media*, *Baja* o *No sensible*. Por ejemplo, un funcionario público de alto rango podría recibir una calificación distinta a la de un trabajador independiente.
- ROS (Reporte de Operación Sospechosa): Indica si el cliente ha sido objeto de un reporte enviado a la Unidad de Información y Análisis Financiero (UIAF), por presentar operaciones inusuales que podrían estar relacionadas con lavado de activos o financiación del terrorismo.

- Medios: Señala si el cliente ha sido mencionado en medios de comunicación definidos por la organización con relación a procesos judiciales, como investigaciones, condenas o extradiciones.
- Riesgo SARLAFT: Calificación basada en el Sistema de Administración del Riesgo de Lavado de Activos y Financiación del Terrorismo (SARLAFT), implementado por la organización como parte de su marco normativo obligatorio.
- RIC (Riesgo Integrado del Cliente): Métrica compuesta que evalúa el riesgo total del cliente tomando en cuenta los cuatro factores definidos por SARLAFT:
    - Clientes: Tipo de cliente (natural o jurídico) y su perfil.
    - Productos: Servicios financieros utilizados (cuentas, créditos, inversiones, seguros, etc.).
    - Canales: Vías de interacción utilizadas, como oficinas físicas, banca digital, cajeros automáticos, entre otros.
    - Jurisdicciones: Zonas geográficas asociadas a la operación del cliente, evaluadas según su nivel de riesgo.
- Requerimientos: Informa si el cliente ha recibido requerimientos formales por parte de entidades judiciales o de control, como la Fiscalía General de la Nación, la Policía Nacional u otras autoridades competentes.
- Relacionados en LC (Listas de Control): Indica si el cliente tiene vínculos familiares (hasta segundo grado de consanguinidad) con personas incluidas en listas de control internas por fraude confirmado.
- Calificación de la Jurisdicción: Evalúa el nivel de riesgo asociado a la ubicación geográfica donde el cliente realiza sus operaciones. Por ejemplo, zonas fronterizas o de difícil acceso pueden implicar un mayor nivel de riesgo.
- Calificación de la Actividad Económica: Nivel de riesgo determinado según la actividad económica que el cliente haya registrado ante la entidad.
- Efectivo: Indica si el cliente ha sido objeto de investigaciones relacionadas con el manejo de efectivo, lo cual puede representar un riesgo adicional para la entidad.

Es importante destacar que, debido a la naturaleza del problema abordado, es completamente esperable encontrar un desbalance tanto en la variable objetivo como en las variables de entrada. Esto se debe a que las etiquetas y clasificaciones utilizadas provienen de procesos altamente específicos y rigurosos dentro de la organización, lo que naturalmente limita la cantidad de casos positivos y puede generar distribuciones poco equilibradas en los datos.

8. ENTENDIMIENTO DE LOS DATOS

En esta etapa se presenta la parte inicial de las actividades realizadas y los resultados obtenidos relacionados con el primer objetivo específico:

“Crear una base de datos anonimizada que recopile el contexto personal de cada cliente, abarcando indicadores de riesgo, su historial transaccional y datos demográficos.”

Dado el compromiso con la confidencialidad exigido por la organización propietaria de los datos, es importante destacar que este apartado no incluye información real de clientes o usuarios.

La recopilación de datos en este proyecto se realizó a partir de cuatro fuentes principales, extraídas de bases de datos relacionales. A continuación, se describe el proceso seguido en esta fase inicial:

1. Recolección de datos:

   Se identificaron cuatro tablas clave dentro del sistema de base de datos relacional de Bancolombia. Estas contenían información esencial sobre los clientes, su comportamiento financiero, transacciones y alertas de fraude:

   - Tabla de clientes: Incluye información personal y demográfica del cliente, como edad, tipo de documento, nacionalidad, entre otros.
   - Tabla corporativa: Contiene información relacionada con el vínculo entre el cliente y las distintas entidades del Grupo Bancolombia.
   - Tabla de riesgo: Reúne variables asociadas a la evaluación del riesgo del cliente, según criterios internos y regulatorios.
   - Tabla transaccional y de modelos: Contiene información relacionada con los modelos de alertas y el comportamiento transaccional del cliente.

2. Integración de fuentes:
   La integración de las tablas se realizó utilizando claves primarias y foráneas para relacionar los datos. El resultado fue un DataFrame estructurado que integra información proveniente de las bases de datos de clientes, corporativa, de riesgo y transaccional, consolidando así variables demográficas, relacionales, de perfil de riesgo y comportamiento transaccional, junto con las etiquetas de fraude asignadas por los expertos. A continuación, se presenta una tabla que contiene las variables del dataset construido, su descripción y los posibles valores que pueden tomar.

Tabla 2. Listado, descripción, posibles valores y tipo de las columnas resultantes

Variable	Descripción	Posibles valores	Tipo de variable (C = Categórica, N = Numérica)
Corporativas			
Id_cliente	Identificación del cliente alertado.	*Número entero*	N
Id_caso	Código único del caso evaluado	*Número entero*	N
recien_vinculado	Indicador de si el cliente fue vinculado recientemente.	*1 (≤ 24 meses), 0 (> 24 meses), -1 (sin información)*	C
segmento	Segmento comercial al que pertenece el cliente	*PERSONAL, PREFERENCIAL, PLUS, ETC.*	C
importador	Indicador de si el cliente realiza actividades económicas de importación.	*1 (Sí), 0 (No)*	C
exportador	Indicador de si el cliente realiza actividades económicas de exportación.	*1 (Sí), 0 (No)*	C
empleado	Indicador de si el cliente es empleado de Bancolombia.	*1 (Sí), 0 (No)*	C
pj_recien_constituida	Persona jurídica constituida hace seis meses o menos	*1 (Sí), 0 (No)*	C
Demográficas			
region	Región geográfica asociada al cliente.	*Antioquia, Caribe, No Registra, Centro, ETC.*	C
tipo_cliente	Indica si es persona natural o jurídica	*Persona Natural, Persona Jurídica*	C
tipo_documento	Tipo de documento de identidad del cliente	*Cédula de Ciudadanía, NIT, Pasaporte, Cédula de Extranjería, ETC.*	C
edad	Edad del cliente	*Número entero*	N
pep	Si el cliente es una persona expuesta políticamente	*1 (Sí), 0 (No)*	C
relacionado_pep	Si el cliente está relacionado con una persona expuesta políticamente	*1 (Sí), 0 (No)*	C
menor	Indica si el cliente tiene menos de 18 años	*1 (Sí), 0 (No)*	C
mayor	Indica si el cliente tiene más de 60 años	*1 (Sí), 0 (No)*	C
Riesgo			

cliente_sensible	Indica si el cliente tiene calificación interna como cliente sensible	*3 (Alta), 2 (Media), 1 (Baja), 0 (No)*	C
ros	Indicador de Reporte de Operación Sospechosa en los últimos 3 años	*0, 1, 2, 3*	C
medios	Si ha sido mencionado en medios con estado procesal condenado o extraditado	*1 (Sí), 0 (No)*	C
riesgo_sarlaft	Riesgo individual del cliente según SARLAFT	*0, 1, 2, 3, 4*	C
requerimientos	Si ha tenido requerimientos de entes legales como policía, fiscalía, etc.	*1 (Sí), 0 (No)*	C
rel_lc	Si tiene relacionados en listas de control vinculadas a alertas de fraude	*1 (Sí), 0 (No)*	C
califica_jur	Calificación de la jurisdicción donde realiza transacciones	*0, 1, 2, 3, 4*	C
califica_act_economica	Medición de riesgo según la actividad económica	*0, 1, 2, 3, 4*	C
ric	Riesgo integrado del cliente según SARLAFT	*0, 1, 2, 3, 4*	C
efectivo	Si ha tenido investigaciones relacionadas con manejo de efectivo	*1 (Sí), 0 (No)*	C
Transaccionales			
supera_perfil_trx	Supera el percentil 99 en el perfil transaccional respecto a su segmento	*1 (Sí), 0 (No), -1 (No registra)*	C
supera_enviados	Supera el percentil 99 en giros enviados respecto a su segmento	*1 (Sí), 0 (No), -1 (No registra)*	C
supera_recibidos	Supera el percentil 99 en giros recibidos respecto a su segmento	*1 (Sí), 0 (No), -1 (No registra)*	C
Variable objetivo			
target	Etiqueta de cliente posiblemente fraudulento	*1 (Fraudulento), 0 (No fraudulento)*	C

3. Anonimización de los datos:

Con el objetivo de garantizar la anonimización de los clientes y proteger su privacidad, se utilizó la función *mask_hash*, integrada en SQL. Esta función fue aplicada a todos los documentos de identificación de los clientes, asegurando que los registros fueran enmascarados de manera efectiva. Este proceso permitió mantener la confidencialidad de los datos sin comprometer su integridad para los análisis posteriores. A continuación, se presenta un ejemplo del enmascaramiento realizado con el ID, acompañado de algunas variables sin enmascarar:

Tabla 3. Ejemplo del enmascaramiento realizado a los documentos

id_cliente	segmento	edad
3f1a8c7b5f4d8e9b9a1c	CONSTRUCTOR	35
7b9c2e1f4a3d6f8e9a0b	CORPORATICO	42
8e9b7c6d5a4f3f1a2b0c	PERSONAS	28

4. Exploración inicial:

El conjunto de datos inicial constaba de un total de 51,727 registros, de los cuales aproximadamente el 10% (5,172 registros) se reservó aleatoriamente como conjunto de evaluación. Estos registros fueron trasladados a un archivo independiente para evitar cualquier contaminación derivada de procesos realizados durante la fase de entrenamiento, como la generación de datos sintéticos. Este conjunto será tratado en detalle durante la etapa de evaluación.

El análisis exploratorio realizado sobre el 90% restante de la base, utilizada para el modelado, permitió identificar aspectos clave sobre la estructura y distribución de los datos, proporcionando información esencial para su comprensión y posterior tratamiento.

1. Historial temporal: Los datos abarcan un período desde el 19 de mayo de 2011 hasta la el 30 de noviembre de 2024.
2. Clientes únicos: El conjunto incluye información de 42,270 clientes únicos, sin registros nulos en el campo de identificación del cliente (*"id_cliente"*).
3. Casos únicos: Se identificaron 46.554 casos únicos, también sin registros nulos en el campo de identificación del caso (*"id_caso"*).
4. Análisis de distribución para variables categóricas: El conjunto de datos cuenta con 26 variables categóricas de entrada y una variable categórica de salida (target). Al realizar el análisis de distribución, se identificó un desbalance en 18 de estas variables. Se consideró como desbalanceada toda variable en la que una sola categoría concentrara más del 80% de los registros. Las variables identificadas fueron: cliente_sensible, ros, medios, requerimientos, rel_lc, tipo_cliente, tipo_documento, pep, relacionado_pep, importador, exportador, efectivo, califica_jur, empleado, pj_recien_constituida, menor, mayor y supera_enviados.

Tras revisar estas variables junto con el equipo experto, se concluyó que dichas distribuciones son coherentes con el comportamiento esperado del negocio. Por tanto, se decidió incluirlas en el desarrollo del modelo, ya que responden a criterios definidos por las áreas funcionales involucradas.

En cuanto a la variable objetivo (target), su distribución es de 73.1% para la clase “no fraude” (0) y 26.9% para la clase “fraude” (1). Aunque esta proporción es relativamente más balanceada que en otros contextos similares, se contempló desde las etapas previas del modelado la necesidad de aplicar técnicas de balanceo para garantizar un mejor desempeño del modelo.

A continuación, se presentan las gráficas de distribución correspondientes a todas las variables categóricas analizadas.

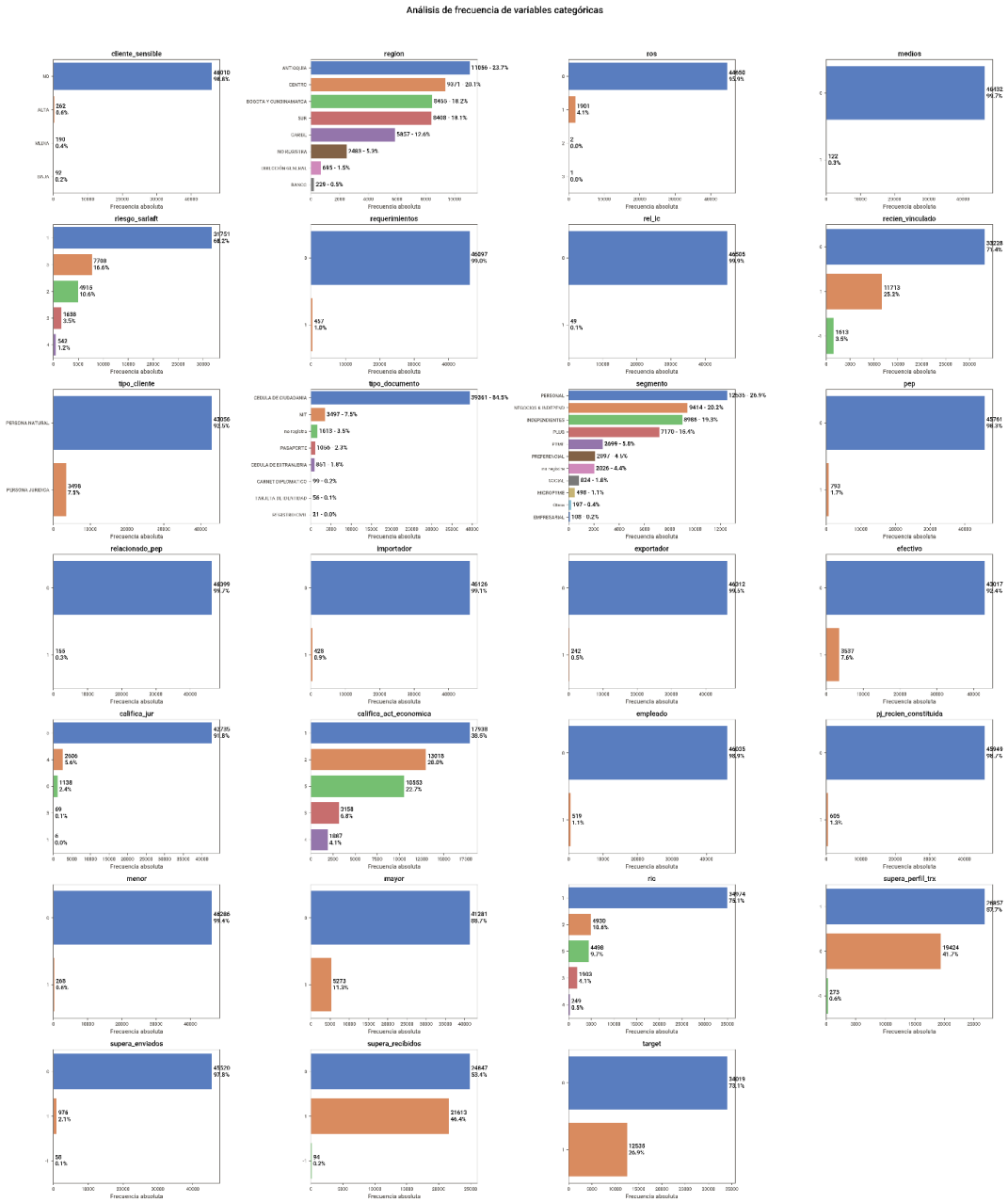
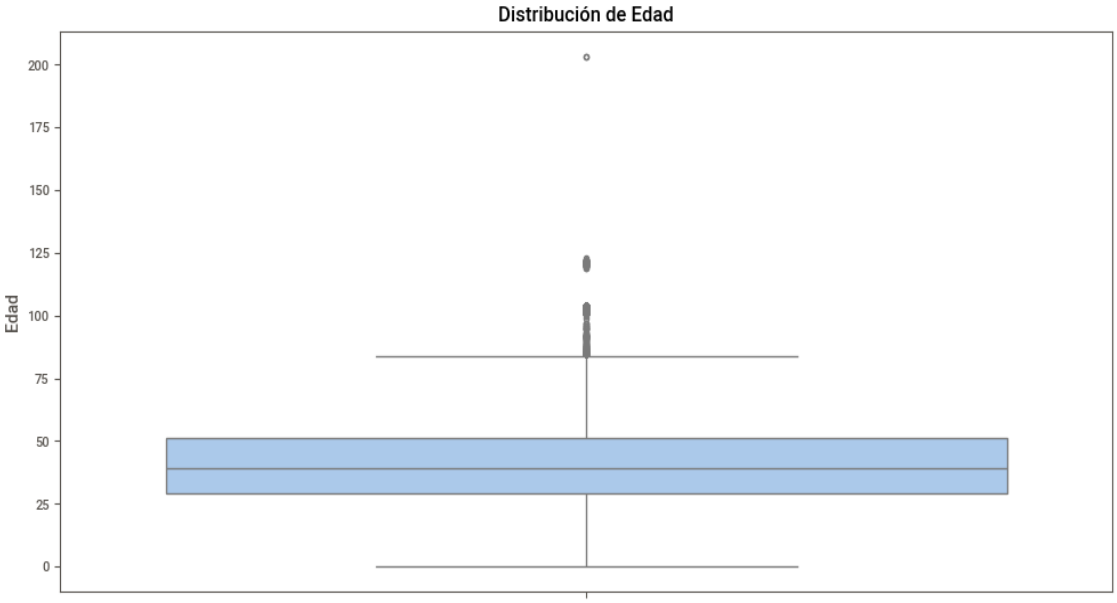


Figura 1. Análisis de frecuencia de las variables categóricas

5. Para este proyecto se contó con una única variable numérica, *Edad*, sobre la cual se realizó un análisis exploratorio específico. Con el fin de evaluar su distribución y posibles valores atípicos, se utilizó un gráfico tipo *boxplot* (diagrama de caja y bigotes), lo que permitió identificar patrones, rangos centrales y extremos que enriquecieron la comprensión del comportamiento de esta variable dentro del contexto del negocio

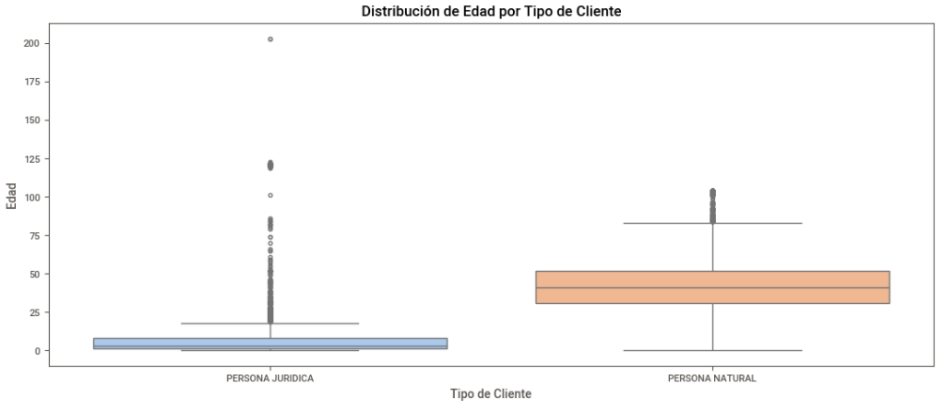


Mínimo	Q1 (25%)	Mediana (Q2)	Q3 (75%)	Máximo	IQR
0.0	29.0	39.0	51.0	203.0	22.0

Figura 2. Boxplot de la variable "Edad"

A partir de la gráfica anterior, se identifican edades que oscilan entre los 0 y los 203 años. Aunque en un primer vistazo estos valores podrían parecer atípicos, durante la etapa de entendimiento del negocio se concluyó, junto con el equipo experto, que estas edades pueden ser válidas dentro de ciertos escenarios específicos según el tipo de cliente.

Para profundizar en este análisis y brindar mayor claridad, a continuación, se presenta una gráfica con un nivel de granularidad superior, segmentando la variable Edad según la categoría de “tipo_cliente”. Esta visualización permite observar cómo se comporta la edad dentro de cada segmento y facilita la identificación de patrones más consistentes con la realidad del negocio.



	Min	Q1	Mediana	Q3	Max	IQR
PERSONA JURIDICA	0.0	1.0	3.0	8.0	203.0	7.0
PERSONA NATURAL	0.0	31.0	41.0	52.0	104.0	21.0

Figura 3. Boxplot de "Edad" por "Tipo de cliente"

La gráfica de la distribución de la variable Edad segmentada por el Tipo de Cliente permitió evidenciar diferencias significativas en su distribución, lo que aportó un contexto valioso para interpretar su comportamiento y, especialmente, para comprender la presencia de valores extremos que, en principio, podrían parecer atípicos.

En el caso de las personas jurídicas, se observó una alta concentración de edades bajas, lo cual resultó coherente con la naturaleza de este tipo de cliente. En estos casos, la “edad” no representaba una edad cronológica, sino el tiempo de constitución o antigüedad de la empresa dentro del sistema financiero. Por esta razón, fue completamente razonable encontrar organizaciones con edades entre 0 y 8 años, lo cual reflejaba nuevas constituciones o registros recientes. Asimismo, se identificaron casos con edades significativamente altas, incluso hasta 203 años. Desde el entendimiento del negocio, estos valores se explicaron como empresas históricas o como resultado de transformaciones societarias, fusiones o absorciones con entidades de larga data. Estos registros fueron revisados junto con el equipo experto y validados como coherentes dentro del contexto del negocio.

Para las personas naturales, la distribución presentó un patrón más esperado, con la mayoría de las edades concentradas entre los 30 y 52 años, lo cual coincidió con el rango típico de edad de la población económicamente activa. También se registraron casos con edades cercanas a 0 años, los cuales se explicaron por la apertura de productos financieros destinados a menores de edad. Estos productos eran gestionados por padres o tutores, como cuentas de ahorro orientadas a la educación futura o programas de ahorro infantil, justificando así su presencia en la base de datos.

Este análisis permitió concluir que los valores extremos en la variable Edad no correspondían a errores o inconsistencias, sino que respondían a realidades específicas del negocio y del tipo de cliente. Por tal motivo, no se aplicaron filtros ni transformaciones adicionales sobre esta variable durante esta etapa previa al modelado, ya que su comportamiento fue considerado válido tras su análisis conjunto con los expertos del equipo de fraude.

6. Análisis multivariado: Dado que, exceptuando la edad, todas las variables son categóricas, se realizó un análisis de asociación mediante la matriz de Theil. Esta matriz está compuesta por coeficientes de asociación asimétricos entre variables categóricas, que cuantifican la proporción de reducción de incertidumbre en una variable dependiente al conocer el valor de una variable independiente. En otras palabras, mide cuánto aporta el conocimiento de una variable para predecir otra, tomando valores entre 0 (sin asociación) y 1 (asociación perfecta).

   Los rangos teóricos definidos para interpretar el coeficiente de Theil's U son los siguientes:

   - 0.00 – 0.10: Asociación muy débil o casi nula
   - 0.10 – 0.30: Asociación débil
   - 0.30 – 0.50: Asociación moderada
   - 0.50 – 0.70: Asociación fuerte
   - 0.70 – 1.00: Asociación muy fuerte

A continuación, se presenta el gráfico detallado, en el cual las columnas representan las variables que influyen sobre las variables en las filas.

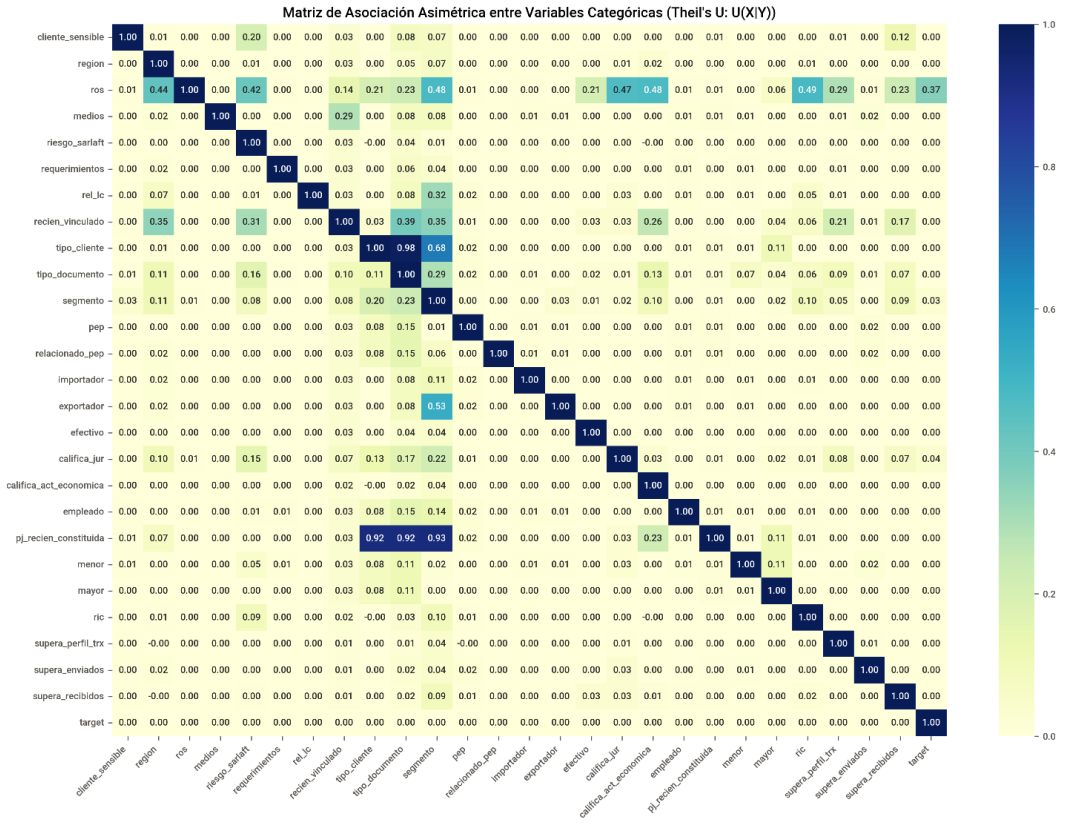


Figura 4. Matriz de asociación asimétrica entre variables categóricas

A continuación, se resumen en una tabla las asociaciones relevantes: Moderadas, fuertes y muy fuertes.

Tabla 4. Asociaciones relevantes, matriz de Theil (U)

Variable 1 (Columna)	Variable 2 (fila)	Valor Theil (U)
tipo_documento	tipo_cliente	0.9817
tipo_cliente	pj_recien_constituida	0.9249
tipo_documento	pj_recien_constituida	0.9249
Segmento	Ros	0.4833
califica_jur	Ros	0.4721
califica_act_economica	Ros	0.4797
supera_perfil_trx	Ros	0.4867
Segmento	recien_vinculado	0.3909
Target	Ros	0.3654

Se identifica una relación muy fuerte entre tipo de documento y tipo de cliente, lo cual es coherente dado que el tipo de documento indica directamente si la persona es natural o jurídica, mientras que el tipo de cliente no brinda esta información de manera tan precisa. Por esta razón,

se concluye que tipo de documento aporta más información relevante, por lo que se decide mantener esta variable y eliminar tipo de cliente, ya que resulta redundante.

Adicionalmente, se observa una relación muy fuerte entre tipo de cliente y pj_recien_constituida, explicada porque la mayoría de las personas naturales (92.5% del total) no son personas jurídicas recién constituidas. Dado que ya se eliminó la variable tipo de cliente, se evita así un posible problema de multicolinealidad en el modelo.

En cuanto a la relación entre tipo de documento y pj_recien_constituida, es similar al caso anterior: el tipo de documento indica si es persona jurídica (por ejemplo, un NIT), pero no especifica si esta fue constituida recientemente o no. Por ello, se decide mantener ambas variables para preservar esta distinción relevante para el análisis.

Finalmente, los valores moderados del coeficiente de Theil (U) se consideran asociaciones interpretables y significativas para el negocio, por lo que no se elimina ninguna variable adicional.

9. PREPARACIÓN DE LOS DATOS

En esta etapa se presenta la parte final de las actividades realizadas y los resultados obtenidos relacionados con el primer objetivo específico:

“Crear una base de datos anonimizada que recopile el contexto personal de cada cliente, abarcando indicadores de riesgo, su historial transaccional y datos demográficos.”

1. Limpieza y preprocesamiento:

   En esta fase se utiliza el análisis exploratorio realizado sobre el conjunto de datos, con el fin de comprender su tipología, distribución, relevancia y posibles inconsistencias. Este diagnóstico permitió establecer criterios fundamentados para la imputación o eliminación de valores faltantes, la eliminación de variables no informativas o redundantes, y la transformación de aquellas que requerían ajustes para su adecuada incorporación en etapas posteriores del análisis y modelado predictivo.

   Durante este proceso se aplicaron las siguientes acciones:

   - Codificación de variables categóricas: Se aplicaron técnicas de one-hot encoding para transformar las variables categóricas en representaciones numéricas que los modelos de aprendizaje automático puedan interpretar. Esta técnica crea una nueva columna para cada categoría presente en las variables categóricas, asignando valores binarios (0 o 1) dependiendo de si una instancia pertenece o no a una categoría específica. Como resultado de este proceso, el conjunto de datos se amplió significativamente, generando un total de 364 variables después de la codificación.

- Escalamiento de variables continuas: Para las variables numéricas, se utilizó la técnica de min-max scaling con el objetivo de normalizar los valores. Este método ajusta las características al rango [0,1], lo que garantiza que todas las variables contribuyan de manera equitativa al modelo y evita que las características con valores más grandes dominen el proceso de aprendizaje.

10. BALANCEO DE LOS DATOS

El objetivo de esta etapa fue:

"Aplicar técnicas de balanceo de datos, asegurando una representación adecuada de las distintas categorías y variables, para evitar sesgos y mejorar la robustez del análisis."

El conjunto de datos posterior al preprocesamiento contenía un total de 46.554 registros, de los cuales 34,031 correspondían a la clase 0 (no fraudulento) y 12,523 a la clase 1 (potencialmente fraudulento), lo que representaba un 26.9% de la clase minoritaria. Este desbalance en la variable objetivo podía afectar negativamente el desempeño de los modelos de clasificación.

Para abordar esta situación, se aplicaron, sobre los datos obtenidos del preprocesamiento, dos técnicas de balanceo con el objetivo de aumentar la cantidad de registros de la clase minoritaria (fraudes). Inicialmente, se definió un incremento del 60% como base para este balanceo. Sin embargo, este porcentaje no era un valor fijo; más adelante se ajustaría dinámicamente dependiendo de los resultados obtenidos durante las pruebas iniciales con los conjuntos de hiperparámetros.

El objetivo de esta estrategia era encontrar un punto de equilibrio adecuado donde el aumento de la clase minoritaria mejorara significativamente el rendimiento del modelo en métricas como el *F1 score* y el *recall*, sin introducir un exceso de datos sintéticos que pudiera generar ruido o afectar la capacidad del modelo para generalizar. Así, el porcentaje final utilizado para el balanceo varió en función de los resultados obtenidos en cada iteración.

Redes Generativas Adversarias (GANs)

Se utilizaron para generar muestras sintéticas de la clase minoritaria. Estas redes fueron entrenadas para aprender la distribución subyacente de los datos reales y generar observaciones artificiales que representaran fielmente el comportamiento de los clientes fraudulentos.

Estructura de la red generadora:

- Capa 1: 128 unidades densas con función de activación ReLU.
- Capa 2: 256 unidades densas con función de activación ReLU.
- Capa de salida: Función de activación Tanh.

Estructura del discriminador:

- Capa 1: 256 unidades densas con función de activación ReLU.
- Capa 2: 128 unidades densas con función de activación ReLU.
- Capa de salida: 1 unidad densa con función de activación Sigmoid.

Se eligieron las funciones de activación en las GANs para optimizar tanto el aprendizaje como la generación de datos. La ReLU fue seleccionada por su eficacia en redes profundas y su capacidad para identificar patrones complejos en datos no lineales, lo que facilita el aprendizaje de las relaciones entre las variables. En el generador, se usó la función Tanh para asegurar que las muestras sintéticas sigan una distribución similar a la de los datos reales, manteniendo su realismo. Por último, la Sigmoid en el discriminador permite distinguir fácilmente entre datos reales y generados, ya que produce probabilidades claras y comprensibles para esta clasificación binaria.

- Synthetic Minority Oversampling Technique (SMOTE)

  SMOTE generó nuevos datos sintéticos mediante interpolaciones entre observaciones cercanas de la clase minoritaria, equilibrando así el conjunto de datos sin eliminar información de la clase mayoritaria.

Luego del balanceo, la clase fraudulenta (clase 1) alcanzó un total de 19,987 observaciones, lo que representa aproximadamente un 37% del total del nuevo dataset (54,018). Esta mejora en la proporción permitió avanzar hacia la etapa de modelado con un conjunto de datos más equilibrado.

Las métricas de desempeño (*precision*, *recall*, *F1-score* y *accuracy*) fueron evaluadas en la etapa de modelado para determinar cuál técnica de balanceo ofrecía mejores resultados en este contexto específico.

11.MODELADO

El objetivo de esta etapa fue “Entrenar un modelo de aprendizaje automático utilizando los datos recopilados de cada cliente”. Para ello, se empleó un enfoque iterativo que permitió identificar las técnicas de modelado más efectivas en función de las características del problema y del dataset disponible.

Tras finalizar las etapas previas, el conjunto de datos se amplió a un total de 364 variables. Este incremento se debe a la aplicación del método One Hot Encoder, que transformó las variables categóricas en una representación numérica, creando una nueva columna para cada categoría

única presente en los datos originales. Debido a la alta dimensionalidad, se decidió implementar una técnica de reducción de dimensiones con el fin de disminuir el costo computacional sin comprometer significativamente la variabilidad original de los datos. Para ello, se optó por aplicar un Análisis de Componentes Principales (PCA). Como parte del proceso, se analizó la varianza explicada en función del número de componentes principales, con el objetivo de identificar el punto óptimo de reducción que conserve la mayor cantidad de información posible.

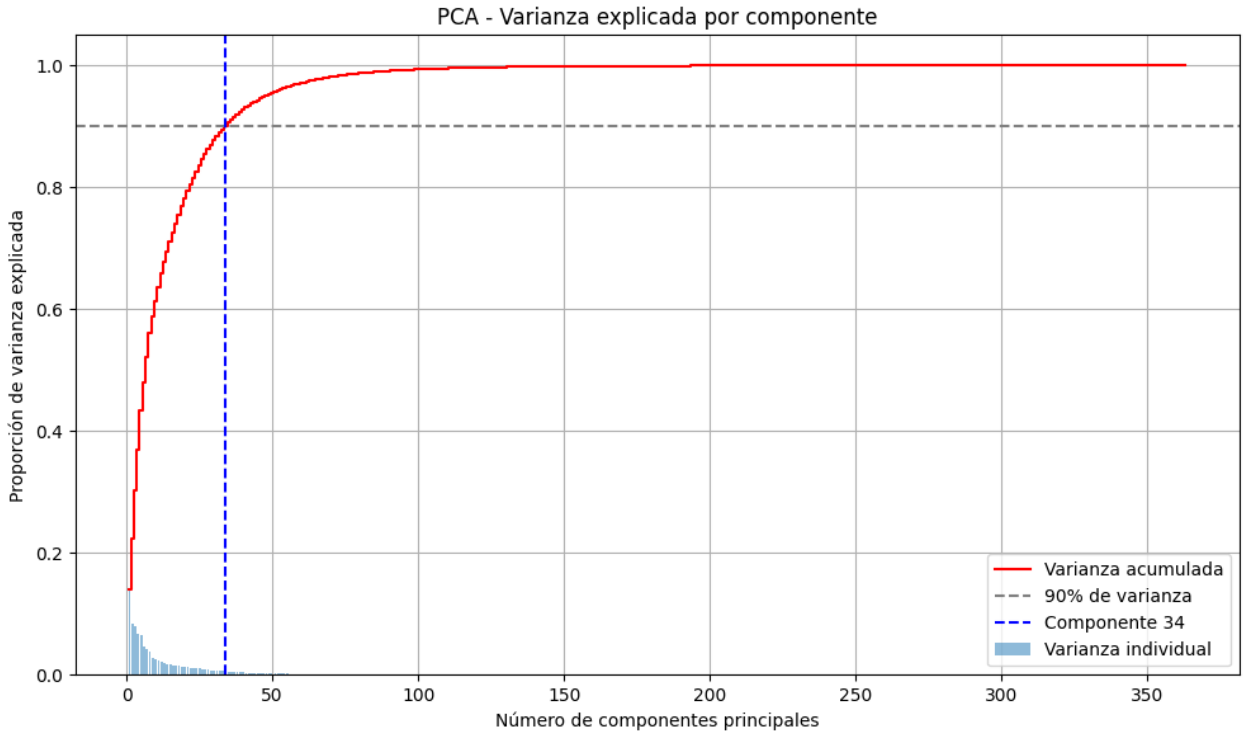


Figura 5.Porcentaje de varianza explicada por número de componentes

Como se muestra en la figura anterior, se identificó que, con 34 componentes es posible explicar aproximadamente el 90% de la varianza total del conjunto de datos. Este umbral se consideró adecuado para garantizar una representación eficiente de la información sin incurrir en pérdida significativa de patrones relevantes.

Por lo tanto, se seleccionaron las primeras 34 componentes principales como base para las etapas posteriores de modelado.

1. Separación en conjuntos de entrenamiento y prueba.

Antes de evaluar los modelos de Machine Learning, el dataset se dividió en dos conjuntos: uno destinado al entrenamiento de los modelos (80% del total de registros) y otro reservado para la validación de los resultados (20%). Esta división garantiza que el modelo se entrene con una

cantidad significativa de datos, mientras se preserva un subconjunto independiente para medir su desempeño real.

Para realizar esta tarea, se utilizó la función train_test_split de la biblioteca Scikit-learn, que asegura una selección aleatoria de los datos y mantiene la proporción entre clases en ambas particiones.

2. Selección inicial de algoritmos.

Los algoritmos seleccionados inicialmente se listan a continuación:

- Random Forest
- Gradient Boosting (XGBoost)
- Decision Tree
- K-Nearest Neighbors (KNN)

La selección de los algoritmos utilizados fue cuidadosamente alineada con la naturaleza del problema de clasificación binaria, cuyo objetivo es identificar clientes fraudulentos. Se priorizaron métodos capaces de manejar datos desbalanceados, especialmente considerando la posibilidad de que la etapa de balanceo no produjera métricas óptimas. Random Forest y Decision Tree fueron elegidos por su habilidad para modelar relaciones no lineales y ofrecer interpretabilidad en las decisiones tomadas. Por otro lado, Gradient Boosting (XGBoost) se eligió por su rendimiento en problemas complejos y su capacidad de optimización. Finalmente, K-Nearest Neighbors (KNN) fue incluido como un enfoque sencillo y efectivo para identificar patrones locales en los datos.

Es importante señalar que los modelos Support Vector Machine (SVM) y Artificial Neural Network (ANN) también fueron evaluados durante la fase exploratoria. Sin embargo, su implementación se vio limitada debido a las características del dataset y las restricciones de recursos de cómputo disponibles (errores por falta de memoria en los equipos). Además, por políticas de confidencialidad, no se permitió el traslado del dataset a sistemas externos al banco, lo que imposibilitó completar su entrenamiento de manera eficiente en el entorno disponible.

3. Entrenamiento de algoritmos.

Los modelos fueron entrenados utilizando los datos balanceados generados en la etapa anterior y evaluados con base en métricas clave como *recall*, *precision*, F1-score y *accuracy*. Estas métricas permitieron medir tanto la capacidad de los modelos para identificar correctamente los clientes fraudulentos como su desempeño general en la clasificación.

A continuación, se presentan los resultados obtenidos por cada modelo, destacando las pruebas realizadas con diferentes configuraciones de hiperparámetros. Este enfoque permitió evaluar las mejoras en el rendimiento los modelos y su capacidad de adaptarse eficazmente al problema planteado.

- K-Nearest Neighbors (KNN)

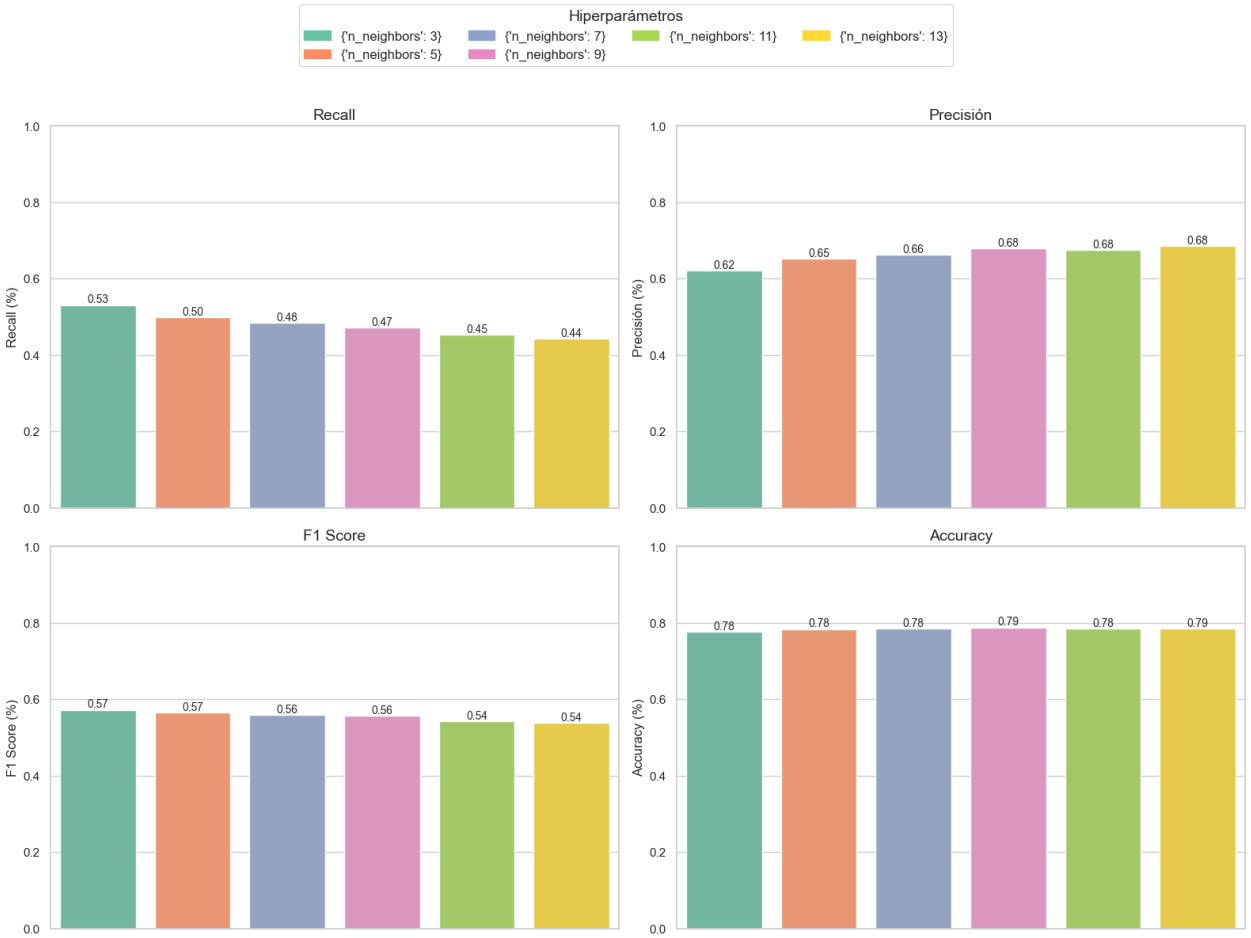


Figura 6. Métricas para el modelo KNN con distintos hiperparámetros

El modelo KNN presentó un desempeño moderado, con un *F1 score* que osciló entre 0.554 y 0.57. Sin embargo, el *recall* se estuvo en un rango de 0.53 y 0.44, lo cual evidenció dificultades significativas del modelo para identificar de manera consistente los casos de fraude. Además, el *accuracy*, que varió entre 78% y 79%, no resulta un indicador confiable en este contexto debido al desbalance de clases, ya que probablemente refleja un buen desempeño únicamente en la clase mayoritaria (no fraude), sin capturar adecuadamente los casos minoritarios.

- Ramdom Forest

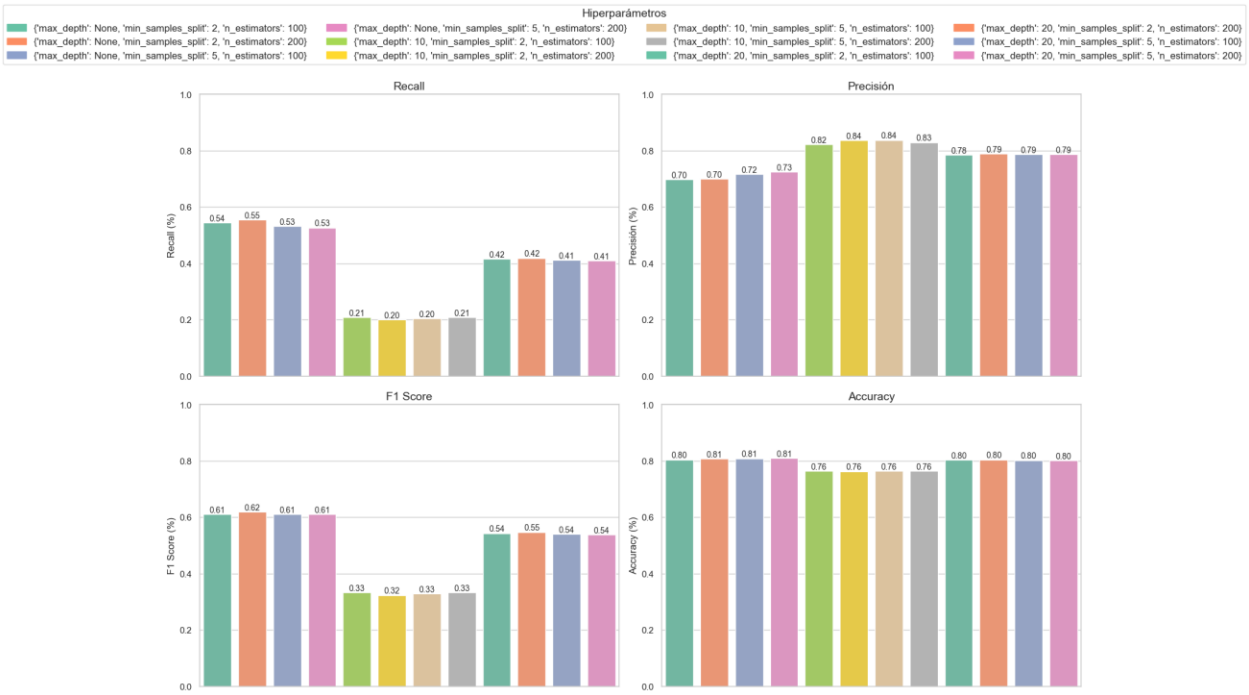


Figura 7.Métricas para el modelo Random Forest con distintos hiperparámetros

El modelo de Random Forest mostró un desempeño variable, con un *F1 score* que alcanzó un valor máximo de 0.62, pero también descendió hasta 0.33. El *recall*, con un pico de 0.55 y un mínimo de 0.21, al igual que KNN, presenta problemas para detectar casos de fraude. Por otro lado, la métrica *precision* fue más sólida, oscilando entre 0.70 y 0.84, lo que sugiere que, cuando el modelo identifica un caso como fraude, es más probable que esté en lo correcto. Finalmente, el *accuracy* tampoco es un indicador confiable debido al desbalance de clases.

- XGBoost

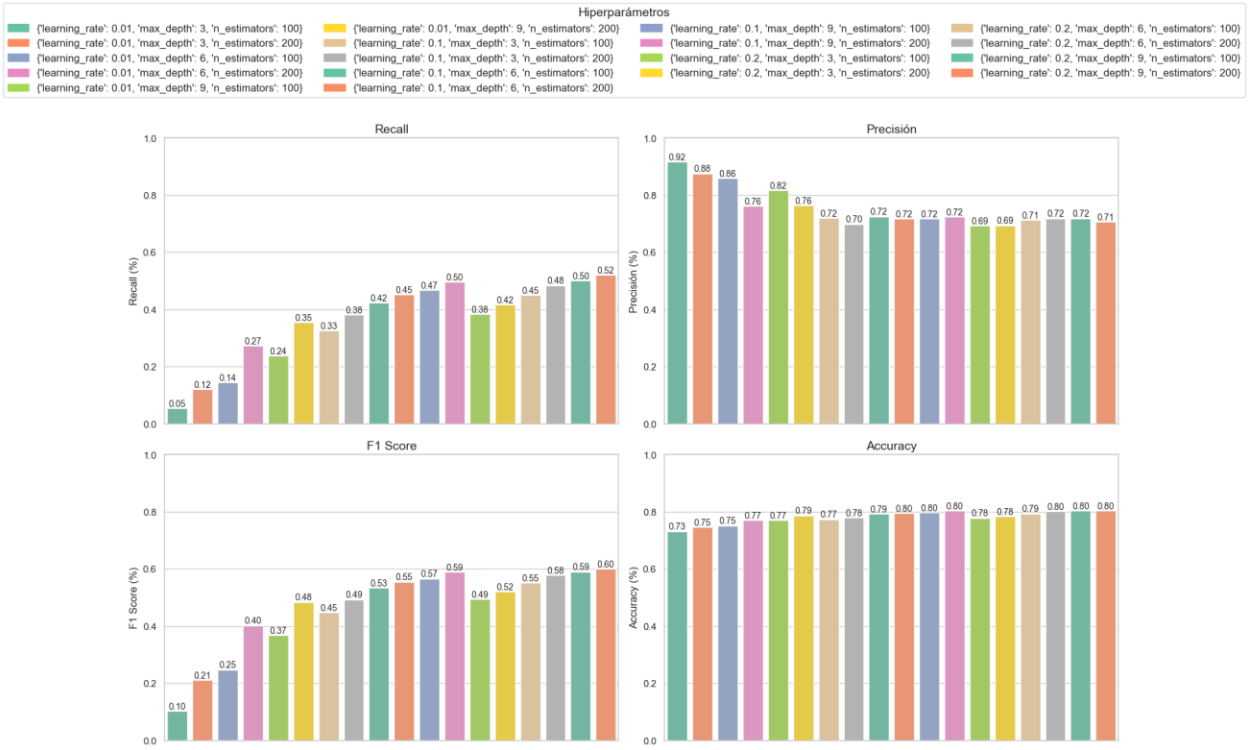


Figura 8. Métricas para el modelo XGBoost con distintos hiperparámetros

El modelo XGBoost alcanzó un valor pico de *F1 score* de 0.60, lo que indica un desempeño moderado en balancear *precision* y *recall*. En este punto máximo, el *recall* fue de 0.52, mostrando una incapacidad para identificar casos de fraude, mientras que *precision* alcanzó un 0.92, reflejando que las predicciones positivas son confiables. Además, el *accuracy* llegó a un 81%, pero, como en los casos anteriores, el desbalance de clases puede afectar mucho esta métrica.

- Decision Tree

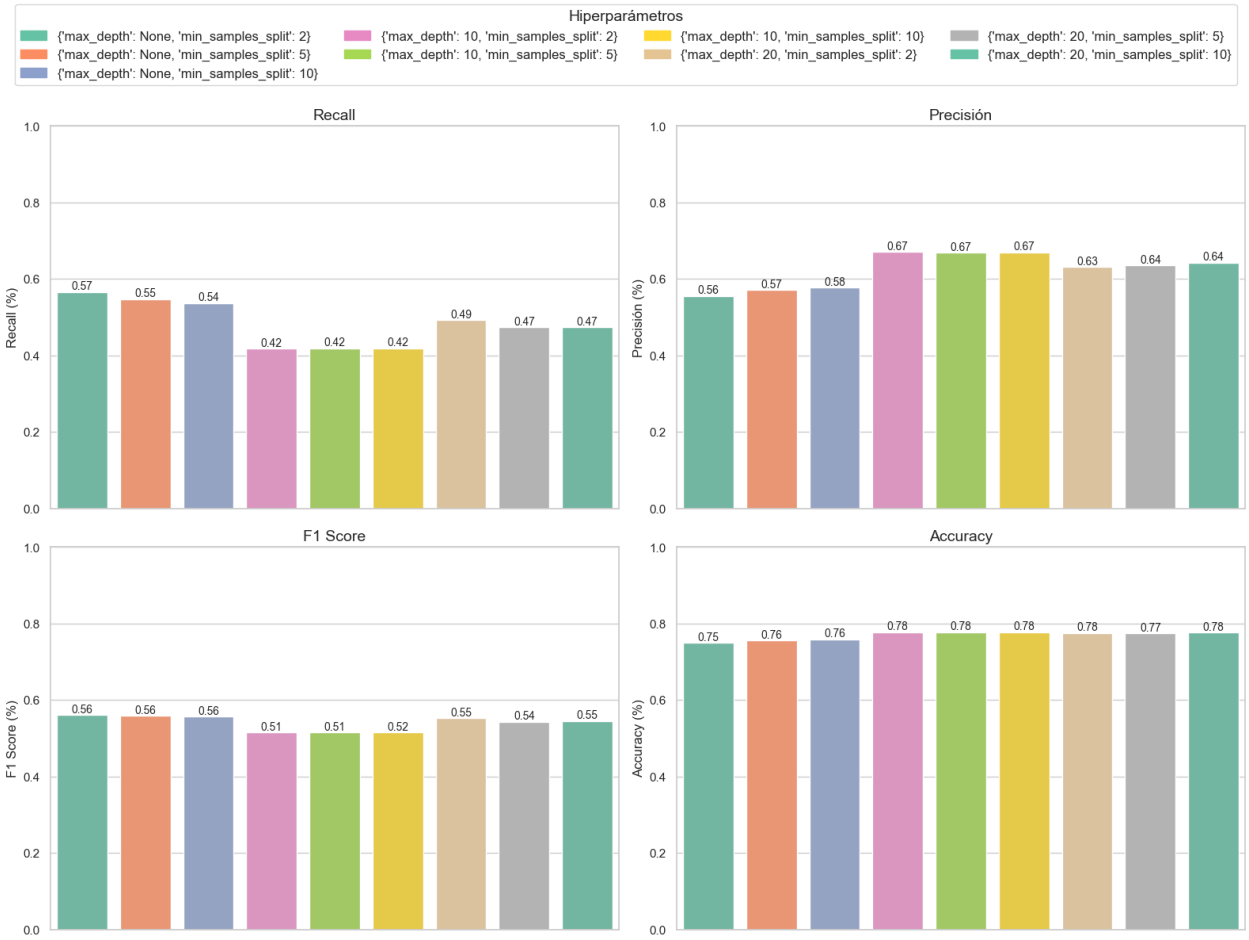


Figura 9. Métricas para el modelo Decision Tree con distintos hiperparámetros

El modelo Decision Tree alcanzó un desempeño moderado en sus mejores resultados, con un *F1 score* máximo de 0.56, lo que refleja un balance aceptable entre *precision* y *recall*. En ese punto, el *recall* llegó a 0.57, indicando una incapacidad para identificar casos positivos, mientras que *precision* fue de 0.67, mostrando que la mayoría de las predicciones positivas fueron correctas. Aún así, los resultados fueron menores que en los otros modelos.

Finalmente, se seleccionó el *F1 score* como la métrica principal de evaluación, dado que proporciona un equilibrio entre *precision* y *recall*. Esta métrica es particularmente adecuada en el contexto del problema, ya que para la compañía resulta igualmente crítico identificar de manera precisa tanto los casos de fraude como aquellos que no lo son.

A continuación, se presenta una tabla que resume los valores de las métricas obtenidas para la configuración óptima de cada modelo.

Tabla 5. Mejores hiperparámetros para cada modelo

Modelo	**Mejores Parámetros**	***Recall***	***Precision***	***F1 score***	***Accuracy***
KNN	{'n_neighbors': 5}	0.53	0.62	0.57	0.78
Random Forest	{'max_depth': None, 'min_samples_split': 2, 'estimators' : 200}	0.55	0.70	0.62	0.904
XGBoost	{'learning_rate': 0.1, 'max_depth': 9,'estimators' : 200}	0.52	0.71	0.60	0.8
Decision Tree	{'max_depth': None, 'min_samples_split': 2}	0.57	0.56	0.56	0.75

Aunque tanto XGBoost como Random Forest mostraron resultados similares en términos de las métricas de evaluación, se optó por continuar el proceso de modelado con XGBoost debido a su capacidad para manejar de manera más eficiente relaciones complejas y no lineales en los datos. Además, XGBoost ofrece mejor escalabilidad en conjuntos de datos más grandes, lo que lo convierte en una opción más robusta para un posterior despliegue en ambiente productivo en la Organización.

Los hiperparámetros iniciales determinados para este modelo son los siguientes:

{'learning_rate': 0.2, 'max_depth': 9, 'n_estimators': 200}

12.EVALUACIÓN

Conjunto de evaluación:

Se cargó el dataframe correspondiente al conjunto de evaluación, aislado previamente durante la etapa de exploración. Este subconjunto equivale al 10% de la base inicial, con un total de 5,172 registros. Los datos pasaron por la etapa de preprocesamiento, que incluyó la aplicación de técnicas de codificación one-hot encoding (OHE) y escalado mediante Min-Max Scaler, para garantizar su compatibilidad con el modelo.

Las métricas de desempeño, como precisión, recall y F1 Score, se calcularon evaluando los modelos entrenados sobre este conjunto de evaluación. Este enfoque aseguró que las métricas reportadas reflejaran el rendimiento del modelo en datos completamente nuevos, no utilizados ni directa ni indirectamente durante el entrenamiento o el ajuste de los hiperparámetros.

Ajuste de hiperparámetros:

Luego de elegir XGBoost como el modelo final, se realizó una iteración más exhaustiva explorando una variedad de hiperparámetros, las dos estrategias de balanceo de clases (SMOTE y GANS) y la técnica de reducción de dimensionalidad PCA. El propósito principal de esta etapa fue maximizar el desempeño del modelo y mejorar su capacidad para identificar casos de fraude con mayor *precision*.

El proceso de ajuste de hiperparámetros se realizó utilizando exclusivamente el conjunto de entrenamiento para entrenar los modelos. Esto garantizó que los modelos se ajustaran únicamente a los datos destinados al aprendizaje, evitando cualquier contaminación del conjunto de evaluación.

Además de emplear las técnicas mencionadas, se experimentó con diferentes porcentajes de balanceo para ajustar la proporción entre la clase mayoritaria y la minoritaria en el conjunto de entrenamiento. Esto permitió analizar cómo variaciones en el nivel de equilibrio afectaban la capacidad del modelo para detectar fraudes sin comprometer su rendimiento en general.

Para el ajuste de hiperparámetros, inicialmente se utilizó la técnica de Random Search como prueba exploratoria. Este método permitió identificar rangos prometedores de valores. Posteriormente, se aplicó una optimización más estructurada mediante Grid Search, que evaluó combinaciones específicas de hiperparámetros dentro de los rangos previamente acotados.

Los rangos considerados para la optimización que fueron entregados por Random Search y aplicados al Grid Search fueron los siguientes:

- colsample_bytree: [0.8, 1]
- learning_rate: [0.01, 0.1, 0.2]
- max_depth: [3, 6, 9, 12, 15]

- n_estimators: [100, 200, 300]
- subsample: [0.8, 1]

A continuación, se presenta una tabla con los resultados obtenidos durante el proceso de optimización, que reflejan el impacto de cada conjunto de hiperparámetros y técnicas de balanceo sobre el desempeño del modelo sobre el conjunto de datos de evaluación.

Tabla 6. Resultados del proceso de optimización de hiperparámetros para XGBoost

% de Balanceo (Fraude vs No Fraude)	Técnica de balanceo	Mejores hiperparámetros	*F1 score* (Métrica principal de evaluación)	*Recall*	*Precision*	*Accuracy*	Tiempo de entrenamiento (min)
27:73	Línea Base	colsample_bytree=1 learning_rate=0.2 max_depth=9 n_estimators=200 subsample=0.8	0.6124	0.5467	0.6961	0.8154	5.6
50:50	SMOTE	colsample_bytree=1 learning_rate=0.2 max_depth=9 n_estimators=200 subsample=0.8	0.6343	0.6147	0.6551	0.8109	9.3
37:67	SMOTE	colsample_bytree=1 learning_rate=0.2 max_depth=9 n_estimators=200 subsample=0.8	0.6343	0.6147	0.6551	0.8109	3.22
37:67	SMOTE + PCA	colsample_bytree=0.8 learning_rate=0.2 max_depth=9 n_estimators=200 subsample=0.8	0.6313	0.5886	0.6806	0.8166	5.34

37:67	SMOTE	colsample_bytree=1 learning_rate=0.2 max_depth=12 n_estimators=300 subsample=0.8	0.6353	0.6055	0.6681	0.8146	5.22
37:67	SMOTE	colsample_bytree=1 learning_rate=0.1 max_depth=15 n_estimators=300 subsample=0.8	0.6374	0.5958	0.6852	0.8192	5.4
37:67	SMOTE + PCA	colsample_bytree=1 learning_rate=0.1 max_depth=15 n_estimators=300 subsample=0.8	0.7463	0.7365	0.7564	0.8168	1.3
50:50	GANS	colsample_bytree=1 learning_rate=0.1 max_depth=9 n_estimators=200 subsample=0.8	0.4296	0.584	0.3146	0.4976	25.4
37:67	GANS	colsample_bytree=1 learning_rate=0.1 max_depth=9 n_estimators=200 subsample=0.8	0.3988	0.6019	0.2982	0.516	18.3

Los resultados de la tabla anterior se presentan de forma gráfica a continuación:

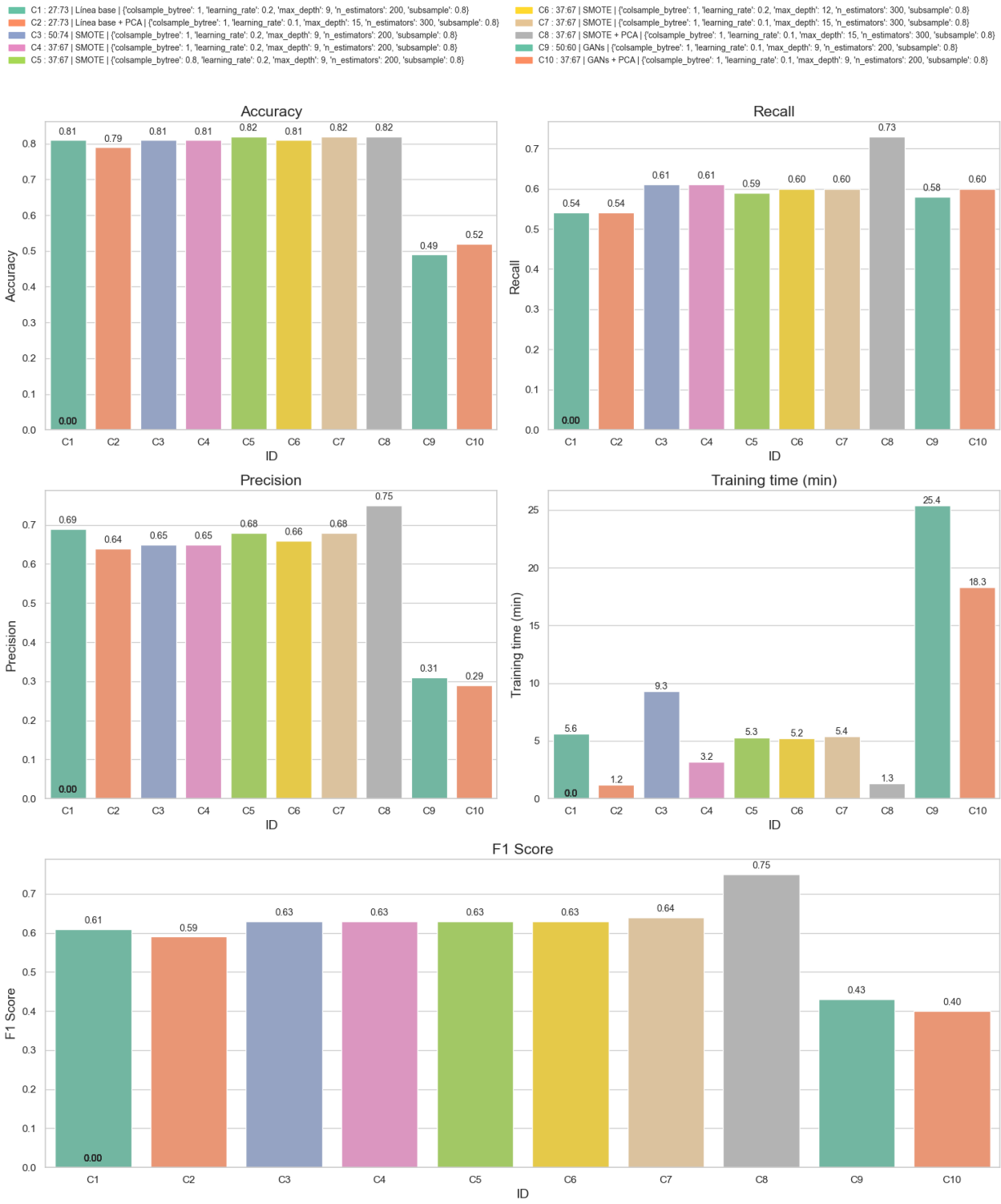


Figura 10. Métricas resultantes del proceso de optimización de hiperparámetros

El modelo más destacado fue el que se aplicó SMOTE con un balanceo 37:67 combinado con PCA. Este modelo logró el mejor *F1 score* (0.7463), acompañado de métricas de *recall* y *precision* equilibradas, además de un tiempo de entrenamiento reducido. Si bien estas métricas representan un avance significativo, no alcanzan niveles óptimos, lo que resalta la necesidad de seguir mejorando el modelo en el futuro. Esto podría incluir la incorporación de nuevas variables explicativas, la implementación de técnicas más avanzadas de preprocesamiento o el uso de algoritmos complementarios que optimicen la capacidad de detección de fraude.

13.CONCLUSIÓN

Este trabajo se enfocó en la aplicación de la ciencia de datos al sector financiero colombiano, específicamente en la mitigación del riesgo asociado al fraude externo. Para ello, se desarrolló un modelo de aprendizaje automático capaz de identificar clientes potencialmente fraudulentos en Bancolombia, integrando variables de riesgo, corporativas, transaccionales y demográficas. El resultado fue un modelo predictivo basado en una técnica de clasificación supervisada, capaz de identificar patrones atípicos con mayor precisión y reducir significativamente la generación de falsos positivos frente a los sistemas de reglas duras.

Este proyecto de grado abordó un desafío importante en el ámbito de la seguridad bancaria: la alta tasa de falsos positivos generada por sistemas tradicionales de monitoreo basados en reglas duras. Estos sistemas, al ignorar el contexto individual de los clientes comprometen la eficiencia operativa. Frente a esta problemática, se diseñó un pipeline que incluyó el tratamiento de datos desbalanceados mediante la técnica de SMOTE, la reducción de dimensionalidad mediante PCA, y una selección y ajuste de modelos supervisados, donde XGBoost emergió como el algoritmo más adecuado para esta problemática.

El modelo final, configurado con un esquema de balanceo 37% (Fraude) y 67% (No fraude) mediante SMOTE combinado con PCA, alcanzó un F1-score de 0.7463, demostrando un desempeño competitivo y equilibrado entre precisión y sensibilidad. Si bien este resultado aún deja margen para optimizaciones, constituye una base sólida para el despliegue de soluciones analíticas más inteligentes y sensibles al contexto en la detección de fraudes.

Además del valor técnico, este proyecto destaca por su articulación entre el frente académico y el entorno corporativo, y por su cuidadoso manejo de los principios de anonimización, confidencialidad y gobernanza de los datos. La colaboración con el equipo de Fraude Externo de Bancolombia permitió un diseño contextualizado, alineado con los procesos reales y con potencial de escalabilidad.

En conclusión, este trabajo demuestra cómo la ciencia de datos puede contribuir no solo al fortalecimiento de la seguridad financiera, sino también a la toma de decisiones estratégicas basadas en evidencia, promoviendo un enfoque más preventivo, preciso y ético en la lucha contra el fraude. El modelo propuesto genera las bases de una herramienta no solo para Bancolombia, sino también como posible referencia para otras entidades del sector financiero que busquen transitar de esquemas tradicionales a soluciones basadas en ciencia de datos.

14.TRABAJO FUTURO

A partir de los hallazgos, logros y limitaciones evidenciadas en el desarrollo de este proyecto, se abren múltiples líneas de investigación y mejora que permitirían fortalecer el alcance y la aplicabilidad del modelo propuesto. Una de las principales proyecciones consiste en su implementación en un entorno productivo dentro del ecosistema tecnológico de Bancolombia. Este despliegue permitiría evaluar su comportamiento bajo condiciones operativas reales, incorporando datos en tiempo real y retroalimentación directa de los equipos encargados del monitoreo. Además, resultará esencial establecer mecanismos de seguimiento continuo que permitan medir su impacto en términos de reducción de falsos positivos, eficiencia en el uso de recursos investigativos y capacidad de detección temprana de eventos sospechosos.

De forma complementaria, se identifica como prioridad el enriquecimiento de las variables y fuentes de información que alimentan el modelo. Si bien el presente trabajo se basó en variables estructuradas históricas, la incorporación de nuevas fuentes de datos, tales como textos no estructurados provenientes de reportes de alerta, menciones en medios de comunicación, historiales judiciales públicos u otras señales contextuales relevantes, podría mejorar significativamente la sensibilidad del modelo ante patrones débiles o dinámicas de fraude emergentes.

Adicionalmente, dada la naturaleza poco frecuente de los casos de fraude y la limitada disponibilidad de etiquetas positivas, se propone explorar enfoques alternativos como el aprendizaje semi-supervisado y no supervisado. Estos métodos podrían identificar perfiles anómalos o patrones atípicos incluso en ausencia de etiquetas explícitas.

Finalmente, resulta pertinente diseñar una estrategia de evaluación integral del impacto organizacional del modelo. Esto incluye no solo métricas cuantitativas sobre su desempeño técnico, sino también su efecto en la eficiencia operativa y los costos asociados a la gestión de alertas. Una evaluación de este tipo podría contribuir a consolidar un marco de referencia útil para otras iniciativas del sector financiero interesadas en aprovechar el potencial de la ciencia de datos para la gestión del riesgo operativo.

15. REFERENCIAS BIBLIOGRÁFICAS

[1] S. Gold, "The evolution of payment card fraud," Computer Fraud & Security, vol. 2014, no. 3, pp. 12–17, 2014. doi: 10.1016/S1361-3723(14)70471-3.

[2] M. M. A.-M. Abu-Orabi and A. F. A. Al Abbadi, "The effects of money laundering on monetary markets introduction," Modern Applied Science, vol. 13, no. 12, pp. 43–51, 2019. doi: 10.5539/mas.v13n12p43.

[3] Superintendencia Financiera de Colombia, "Consideraciones generales," Circular Externa 041, junio de 2007, p. [1]. Disponible en: https://www.fasecolda.com/cms/wp content/uploads/2019/08/ce041-2007-anexo.pdf.

[4] Superintendencia Financiera de Colombia, "2.1. Riesgo Operativo (RO)," Circular Externa 041, junio de 2007, p. 1. Disponible en: https://www.fasecolda.com/cms/wp content/uploads/2019/08/ce041-2007-anexo.pdf.

[5] Superintendencia Financiera de Colombia, "2.6.1.2. Fraude Externo," Circular Externa 041, junio de 2007, p. 2. Disponible en: https://www.fasecolda.com/cms/wp content/uploads/2019/08/ce041-2007-anexo.pdf.

[6] J. C. Peralta, Sistema de administración de riesgo de lavado de activos y financiación del terrorismo (SARLAFT) adecuado para Seaboard Overseas Colombia. [Online]. Available: http://hdl.handle.net/10554/54041.

[7] H. F. M. Morales, "La importancia del cumplimiento de la obligación del Reporte de Operación Sospechosa (ROS) de blanqueo de capitales algunas consideraciones generales: The importance of compliance with the obligation of the Suspicious Operation Report (ROS) of money laundering some general considerations," Constructos Criminológicos, vol. 3, no. 4, pp. 41-62, 2023, doi: 10.29105/cc3.4-42.

[8] E. A. Duque-Grisales, J. Molina Flórez, y N. Ossa Núñez, "Operación del sistema de autocontrol y gestión del riesgo de lavado de activos y financiación del terrorismo en empresas del sector comercial," Revista CINTEX, vol. 23, no. 1, pp. 32–42, 2018, doi: 10.33131/24222208.306.

[9] F. W. Gutiérrez Zanelli, "Amenazas y vulnerabilidades de la regulación de las Personas Expuestas Políticamente (PEP) aplicable a los sujetos obligados supervisados por la Superintendencia de Banca, Seguros y AFP," 2019.

[10] J. A. Restrepo Rodríguez y C. F. Linares, "Principio de equidad establecido en el artículo 363 de la constitución política de Colombia en tributación aplicada a personas naturales y jurídicas," 2018. [En línea]. Disponible en: https://hdl.handle.net/20.500.12494/6001.

[11] J. L. Harrington, Relational Database Design and Implementation. Burlington, MA: Morgan Kaufmann, 2016.

[12] E. Godoc, SQL: Los fundamentos del lenguaje. Francia: Ediciones ENI, 2014.

[13] J. L. García Garmendia and F. Maroto Monserrat, "Interpretación de resultados estadísticos," Medicina Intensiva, vol. 42, no. 6, pp. 370–379, 2018. doi: 10.1016/j.medin.2018.03.006.

[14] F. Bliemel, "Theil's Forecast Accuracy Coefficient: A Clarification," Journal of Marketing Research, vol. 10, no. 4, pp. 444–446, Nov. 1973.

[15] P. Rodríguez, M. A. Bautista, J. Gonzalez, and S. Escalera, "Beyond one-hot encoding: Lower dimensional target embedding," Image and Vision Computing, vol. 75, pp. 21–31, 2018.

[16] K. Mahalakshmi and P. Sujatha, "The role of exploratory data analysis and pre-processing in the machine learning predictive model for heart disease," in 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), pp. 1–8, May 2023

[17] J. E. Martinelli, Clasificación de datos desbalanceados. La Plata, Argentina: Universidad Nacional de La Plata, 2022. Tesis de grado. [En línea]. Disponible en: https://sedici.unlp.edu.ar/handle/10915/147410

[18] R. Qaddoura and M. M. Biltawi, "Improving Fraud Detection in An Imbalanced Class Distribution Using Different Oversampling Techniques," 2022 International Engineering Conference on Electrical, Energy, and Artificial Intelligence (EICEEAI), Zarqa, Jordan, 2022, pp. 1 5, doi: 10.1109/EICEEAI56378.2022.10050500

[19] E. Strelcenia and S. Prakoonwit, "Generating Synthetic Data for Credit Card Fraud Detection Using GANs," 2022 International Conference on Computers and Artificial Intelligence Technologies (CAIT), Quzhou, China, 2022, pp. 42-47, doi: 10.1109/CAIT56099.2022.10072179

[20] A. D. Jiménez Alfaro y J. V. Díaz Ospina, "Revisión sistemática de literatura: Técnicas de aprendizaje automático (Machine Learning)," Cuaderno Activa, vol. 13, no. 1, pp. 113–121, 2022. [En línea]. Disponible en: https://ojs.tdea.edu.co/index.php/cuadernoactiva/article/view/849.

[21] S. Roy, P. Saini, T. H. Thakkar, R. Maranan, A. R. Salve and K. Hemabala, "Data Mining Technology In Student Management Using Svm Techniques," 2023 6th International Conference on Contemporary Computing and Informatics (IC3I), Gautam Buddha Nagar, India, 2023, pp. 1449-1453, doi: 10.1109/IC3I59117.2023.10398099.

[22] N. S. S. Pranavi, T. K. S. S. Sruthi, B. J. Naga Sirisha, M. S. Nayak and V. S. Gupta Thadikemalla, "Credit Card Fraud Detection Using Minority Oversampling and Random Forest Technique," 2022 3rd International Conference for Emerging Technology (INCET), Belgaum, India, 2022, pp. 1-6, doi: 10.1109/INCET54531.2022.9824146.

[23] S. Poojitha and K. Malathi, "An Original Approach to Identify the Better Accuracy in Credit Card Fraud Transaction by Comparing Logistic Regression with K-Nearest Neighbours Algorithm," 2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, 2022, pp. 6-11, doi: 10.1109/ICTACS56270.2022.9987804.

[24] S. Lei, K. Xu, Y. Huang and X. Sha, "An XGBoost based system for financial fraud detection," E3S Web of Conferences, vol. 214, p. 02042, 2020. doi: 10.1051/e3sconf/202021402042.

[25] J. Su and H. Zhang, "A fast decision tree learning algorithm," in Proceedings of the 21st AAAI Conference on Artificial Intelligence (AAAI), vol. 6, pp. 500–505, 2006.

[26] J. M. Navarro Céspedes, G. M. Casas Cardoso, and E. González Rodríguez, "Análisis de componentes principales y análisis de regresión para datos categóricos. Aplicación en la hipertensión arterial," Revista De Matemática: Teoría Y Aplicaciones, vol. 17, no. 2, pp. 199–230, 2010. doi: 10.15517/rmta.v17i2.2128

[27] P. Kaelo and M. M. Ali, "Some Variants of the Controlled Random Search Algorithm for Global Optimization," Journal of Optimization Theory and Applications, vol. 130, pp. 253–264, 2006. doi: 10.1007/s10957-006-9101-0

[28] P. Liashchynskyi and P. Liashchynskyi, "Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS," CoRR, vol. abs/1912.06059, 2019. [Online]. Available: http://arxiv.org/abs/1912.06059

[29] C. Goutte and E. Gaussier, "A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation," in Advances in Information Retrieval, D. E. Losada and J. M. Fernández-Luna, Eds., ECIR 2005, Lecture Notes in Computer Science, vol. 3408, Springer, Berlin, Heidelberg, 2005. [Online]. Available: https://doi.org/10.1007/978-3-540-31865-1_25

[30] R. Davila Moran, R. Castillo-Sáenz, A. Vargas-Murillo, L. Dávila, E. García-Huamantumba, C. García-Huamantumba, R. Cajas, y C. Guanilo, "Aplicación de Modelos de Aprendizaje

Automático en la Detección de Fraudes en Transacciones Financieras," Data and Metadata, vol. 2, p. 109, 29-Oct-2023, Disponible: https://doi: 10.56294/dm2023109.

[31] W. Hilal, S. A. Gadsden, and J. Yawney, "Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances," *Expert Systems with Applications*, vol. 193, p. 116429, May 2022. [Online]. Disponible: https://doi.org/10.1016/j.eswa.2021.116429.

[32] N. S. S. Pranavi, T. K. S. S. Sruthi, B. J. Naga Sirisha, M. S. Nayak, and V. S. Gupta Thadikemalla, "Credit Card Fraud Detection Using Minority Oversampling and Random Forest Technique," in 2022 3rd International Conference for Emerging Technology (INCET), Belgaum, India, 2022, pp. 1-6, doi: 10.1109/INCET54531.2022.9824146.

[33] W. Deng, Z. Huang, J. Zhang and J. Xu, "A Data Mining Based System For Transaction Fraud Detection," 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), Guangzhou, China, 2021, pp. 542-545, doi: 10.1109/ICCECE51280.2021.9342376.

[34] Y. K. Saheed, M. A. Hambali, M. O. Arowolo and Y. A. Olasupo, "Application of GA Feature Selection on Naive Bayes, Random Forest and SVM for Credit Card Fraud Detection,"2020 International Conference on Decision Aid Sciences and Application (DASA), Sakheer, Bahrain, 20 2020, pp. 1091-1097, doi: 10.1109/DASA51403.2020.9317228.

[35] J. S. Saltz, "CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps," 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 2021, pp. 2337 2344, doi: 10.1109/BigData52589.2021.9671634.