ESCUELA DE NEGOCIOS

MAESTRÍA EN INTELIGENCIA DE NEGOCIOS Y CIENCIA DE DATOS

PREDICCIÓN DE FRAUDE FINANCIERO UTILIZANDO TÉCNICAS DE MACHINE LEARNING PARA UNA INSTITUCIÓN FINANCIERA

Profesor
MANUEL EUGENIO MOROCHO CAYAMCELA

Autor
DIEGO OMAR ORTIZ RUIZ

2024

RESUMEN

El fraude financiero es un problema creciente que afecta a instituciones financieras y empresas en todo el mundo, causando pérdidas significativas y erosionando la confianza de los clientes. Este proyecto se centra en el desarrollo de un sistema de predicción de fraude financiero utilizando técnicas avanzadas de Machine Learning. El objetivo principal es identificar transacciones fraudulentas de manera precisa y eficiente, minimizando los falsos positivos y maximizando la detección de actividades fraudulentas.

Para alcanzar este objetivo, se ha seguido una metodología que incluye la recopilación y preprocesamiento de datos, selección de características relevantes, y la implementación y evaluación de varios algoritmos de Machine Learning, como Regresión Logística, Árboles de decisión y Random Forest. El desempeño de cada modelo se ha evaluado utilizando métricas como la precisión, la sensibilidad, la especificidad y el área bajo la curva ROC (AUC-ROC).

Los resultados obtenidos demuestran que los modelos de Machine Learning pueden detectar patrones complejos y sutiles en los datos que son indicativos de fraude, proporcionando una herramienta poderosa para las instituciones financieras. Además, este enfoque puede ser adaptado y escalado para diferentes tipos de fraudes y sectores, lo que lo convierte en una solución versátil y eficaz.

En conclusión, la implementación de técnicas de Machine Learning para la predicción de fraude financiero ofrece una mejora significativa en la capacidad de las organizaciones para protegerse contra el fraude, reduciendo pérdidas económicas y fortaleciendo la confianza de los clientes.

ABSTRACT

Financial fraud is a growing problem affecting financial institutions and businesses worldwide, causing significant losses and eroding customer trust. This project focuses on developing a financial fraud prediction system using advanced Machine Learning techniques. The main objective is to identify fraudulent transactions accurately and efficiently, minimizing false positives and maximizing the detection of fraudulent activities.

To achieve this goal, a methodology has been followed that includes data collection and preprocessing, selection of relevant features, and the implementation and evaluation of several Machine Learning algorithms, such as Logistic Regression, Decision Trees, and Random Forest. The performance of each model has been evaluated using metrics such as accuracy, sensitivity, specificity, and area under the ROC curve (AUC-ROC).

The results obtained demonstrate that Machine Learning models can detect complex and subtle patterns in data that are indicative of fraud, providing a powerful tool for financial institutions. Furthermore, this approach can be adapted and scaled for different types of fraud and sectors, making it a versatile and effective solution.

In conclusion, implementing Machine Learning techniques for financial fraud prediction offers a significant improvement in the ability of organizations to protect themselves against fraud, reducing financial losses and strengthening customer confidence.

ÍNDICE DEL CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

El fraude financiero es un problema significativo y en aumento que enfrenta el sector financiero global. Con la creciente digitalización de las transacciones financieras, las técnicas tradicionales de detección de fraude, como la verificación manual, han demostrado ser ineficaces, costosas y lentas. La evolución de la inteligencia artificial y, en particular, del Machine Learning (ML), ha permitido desarrollar métodos más sofisticados y eficientes para detectar actividades fraudulentas en tiempo real.

El uso de algoritmos de ML para la detección de fraude financiero se basa en la capacidad de estos algoritmos para aprender de grandes volúmenes de datos y adaptarse a patrones nuevos y complejos. Estos modelos pueden identificar transacciones anómalas que podrían indicar actividades fraudulentas, mejorando significativamente la precisión y reduciendo los falsos positivos en comparación con los métodos tradicionales. Según un estudio reciente, los algoritmos de ML, como las redes neuronales artificiales (ANN) y las máquinas de vectores de soporte (SVM), son especialmente efectivos para la detección de fraudes, con aplicaciones predominantes en la detección de fraudes como tarjetas de crédito (Ali et al., 2022).

En este proyecto, se implementarán y evaluarán varios modelos de ML para la predicción de fraude financiero. La metodología incluye la recopilación y preprocesamiento de datos transaccionales reales, la selección de características relevantes y la evaluación de diversos algoritmos de ML. Los modelos considerados incluyen la regresión logística, los árboles de decisión, Random Forest y las redes neuronales. Cada modelo será evaluado en términos de precisión, sensibilidad, especificidad y área bajo la curva ROC (AUC-ROC).

Además, se abordarán aspectos éticos y de privacidad relacionados con la recopilación y el uso de datos financieros. La protección de la información del

cliente y el cumplimiento de las normativas vigentes son esenciales para asegurar la integridad y legalidad del sistema de detección de fraude.

La implementación de técnicas de ML en la detección de fraude financiero no solo mejora la capacidad de las instituciones para identificar y prevenir actividades fraudulentas, sino que también contribuye a reducir pérdidas económicas y a fortalecer la confianza de los clientes. Este proyecto tiene como objetivo demostrar la eficacia de estos modelos y proporcionar una base sólida para futuros desarrollos en el campo de la detección de fraudes financieros.

REVISIÓN DE LITERATURA

En esta sección se proporciona una visión general de la literatura más relevante y relacionada con el presente estudio.

1. Investigación del impacto beneficioso del modelado basado en segmentación para la calificación crediticia.

   De acuerdo con (Idbenjra et al., 2024) la investigación evalúa el impacto beneficioso del modelado basado en segmentación, comparando el modelo de hoja logit (LLM), que se basa en regresión logística (LR) y árboles de decisión. Utilizando un extenso conjunto de datos de calificación crediticia con 65,536 clientes activos, el estudio se centró en tres métricas de evaluación: AUC, elevación del decil superior y ganancias. En resumen, el LLM destaca como clasificador para decisiones crediticias debido a su capacidad para combinar un rendimiento predictivo sólido con conocimientos interpretables, lo que puede informar decisiones gerenciales.
2. Un modelo de calificación crediticia de dos etapas basado en un Random Forest: Evidencia de pequeñas empresas chinas.

   Según (Zhou et al., 2023), la calificación crediticia de las pequeñas empresas es compleja y costosa, lo que la convierte en un campo de investigación desafiante. Utilizando datos de préstamos de 3,045 pequeñas empresas en China, se diseñó un sistema experto de dos etapas para la predicción de incumplimiento, que cuantifica las variables y umbrales clave. Primero, se aplicó SMOTE para manejar los datos desequilibrados, y luego se empleó un modelo de Random Forest para crear funciones crediticias predictivas. Las pruebas de solidez muestran que la metodología propuesta supera a otros modelos de aprendizaje automático, y estos resultados son consistentes con observaciones en otros países.
3. Investigación sobre el Modelo de Predicción Predeterminada del Riesgo de Crédito Corporativo Basado en el Algoritmo de Análisis de Big Data.

(Xianyu & Hai, 2023) Establece un modelo de predicción de incumplimiento para el riesgo crediticio corporativo, optimizando diferentes modelos y comparándolos en función de su rendimiento, y analizando la robustez del modelo óptimo. El conjunto de datos utilizado proviene de bases de datos como CSMAR, Choice Data, Dongfang Fortune, Sina Financial y Tonghua Shun, e incluye 21 índices financieros y no financieros de más de 1,000 empresas cotizadas. Estos datos fueron estandarizados, equilibrados y normalizados, y se utilizó el coeficiente de correlación para filtrar los índices.

Se desarrollaron dos modelos de aprendizaje profundo, una red neuronal convolucional y una red neuronal recurrente, basados en el marco PyTorch en la plataforma Spark. Estos modelos se compararon con dos modelos tradicionales de aprendizaje automático: Random Forest y una regresión logística. Los resultados experimentales mostraron que la red neuronal recurrente es el modelo óptimo, con una tasa de precisión de 0.93, una tasa de recuperación de 0.96 y un valor F1 de 0.93.

Para el modelo óptimo de red neuronal recurrente, se analizó la robustez modificando el número de indicadores, cambiando el número de muestras y eliminando factores no financieros. Los resultados indicaron que los indicadores de evaluación del modelo no varían significativamente bajo estas condiciones, demostrando así una buena robustez del modelo.

4. Gestión y análisis de riesgos financieros basados en datos a gran escala utilizando estrategias de aprendizaje automático.

De acuerdo con (Murugan & T, 2023), esta investigación analizó y procesó conjuntos de datos a gran escala antes del entrenamiento y los evaluó utilizando tres modelos: K-vecino más cercano (KNN) basado en clústeres, regresión logística (LR) basada en clústeres y XGBoost basado en clústeres, por su capacidad para predecir incumplimientos de préstamos y su probabilidad de ocurrencia. Los resultados de la simulación del modelo propuesto demostraron un rendimiento superior en la evaluación de riesgos financieros con datos a gran escala en comparación con métodos de última generación. Para este estudio, se

utilizó un conjunto de datos en línea de los últimos años. Como resultado, XGBoost y K-vecino más cercano (KNN) son recomendados para la gestión de riesgos financieros con la implementación de IoT.

5. Regresion logistica v/s Arboles de decisión en el riesgo crediticio.

   El objetivo de (Segoviano, 2023), fue comparar el rendimiento de dos algoritmos de aprendizaje supervisado, Árbol de Decisión y Regresión Logística, en la predicción del riesgo crediticio. En este contexto, la regresión logística se utiliza para analizar y predecir el nivel de riesgo asociado con un prestatario específico, recopilando datos relevantes como historial crediticio, ingresos y nivel de endeudamiento. Esta técnica identifica patrones y relaciones entre estas variables.

   Por otro lado, los árboles de decisión son una técnica de modelado utilizada para tomar decisiones basadas en múltiples variables. Son interpretables y proporcionan una comprensión clara del razonamiento detrás de las decisiones.

   En este estudio, se evaluó la eficiencia de los modelos de regresión logística y árboles de decisión para predecir el riesgo crediticio. La regresión logística mostró una mayor eficiencia con una precisión de 0.93, mientras que los árboles de decisión alcanzaron una precisión de 0.83 al entrenar ambos modelos con el mismo conjunto de datos.

6. Aprendizaje automático para la puntuación de crédito: Mejora de la regresión logística con efectos de árbol de decisión no lineales.

   Según (Dumitrescu et al., 2022), propone un método de calificación crediticia interpretable y de alto rendimiento llamado regresión de árbol logístico penalizado (PLTR), que utiliza información de árboles de decisión para mejorar el rendimiento de la regresión logística. PLTR permite capturar efectos no lineales presentes en los datos de calificación crediticia, manteniendo al mismo tiempo la interpretabilidad del modelo de regresión logística. Las simulaciones de Monte Carlo y las aplicaciones empíricas que utilizan cuatro conjuntos de datos reales de incumplimiento crediticio demuestran que PLTR predice el riesgo crediticio con mayor

precisión que la regresión logística y se compara competitivamente con el método Random Forest.

7. Modelo y aplicaciones de calificación de calidad crediticia de los prestatarios, con análisis discriminante de incumplimiento basado en la máquina de aprendizaje extremo.

   A través de la evaluación del método del peso de la evidencia y el cálculo del valor de la información (IV). (Pang et al., 2021), propone un método para evaluar las cualidades crediticias de los prestatarios utilizando la máquina de aprendizaje extremo, el algoritmo difuso c-means (FCM) y el cálculo de una matriz de confusión. Mediante la selección de índices de calificación crediticia, se estableció un modelo de calificación crediticia del prestatario. Además, se diseñaron algoritmos específicos para la selección de índices de calificación y para la calificación de la calidad crediticia.

   El estudio recopiló datos de 7,706 prestatarios de préstamos de Renren, una plataforma de préstamos en Internet. Se calcularon las puntuaciones crediticias, la probabilidad de incumplimiento y la tasa de pérdida por incumplimiento de cada tipo de prestatario, y se analizó el estado de pago de los prestatarios. Los resultados experimentales muestran que la precisión general del modelo de calificación crediticia es del 98.5%, con una precisión del 98.9% para las muestras no morosas y del 88.3% para las muestras morosas. La alta precisión del modelo demuestra su potencial para proporcionar valores de referencia importantes y orientación científica para bancos, instituciones financieras y principales plataformas financieras.

8. Modelo de riesgo crediticio basado en datos del registro de crédito del Banco Central.

   Según (Doko et al., 2021), las técnicas de ciencia de datos y aprendizaje automático ayudan a los bancos a optimizar operaciones, mejorar análisis de riesgos y obtener ventajas competitivas. Este estudio evalúa diferentes modelos de aprendizaje automático para crear un modelo preciso de evaluación del riesgo crediticio, utilizando datos del registro de crédito real

del Banco Central de la República de Macedonia del Norte. Se compararon cinco modelos: regresión logística, árbol de decisión, Random Forest, máquinas de vectores de soporte (SVM) y redes neuronales. Los mejores modelos según la puntuación F1 fueron el árbol de decisión, Random Forest y regresión logística.

9. Modelo de clasificación de riesgo crediticio utilizando Random Forest en financiera del Ecuador

Para (Freire López, 2021), este estudio propone crear un modelo basado en un algoritmo de inteligencia artificial para clasificar a los clientes de una organización como "buenos" o "malos" pagadores, en función de diversas variables consideradas importantes para el análisis. Utilizando la base de datos de clientes de crédito de Insotec, que abarca desde 2017 hasta 2020, con 18 variables y 63,896 registros, se implementó la metodología CRISP-DM para el desarrollo del modelo de clasificación.

El modelo, implementado con Random Forest, alcanzó una precisión superior al 97%, con un porcentaje de error del 2.8%, incluyendo un 2.1% de falsos positivos y un 11.1% de falsos negativos. Esta herramienta permite clasificar automáticamente a los clientes, facilitando la concesión de créditos de manera más rápida y con menor riesgo.

10. Predicción de incumplimiento de préstamos mediante árboles de decisión y Random Forest: un estudio comparativo.

Para (Madaan et al., 2021), uno de los principales desafíos que enfrenta el sector bancario en esta economía en constante cambio es la creciente tasa de incumplimiento de préstamos. A las autoridades bancarias les resulta cada vez más difícil evaluar correctamente las solicitudes de préstamos y gestionar los riesgos de incumplimiento. Este estudio propone dos modelos de aprendizaje automático para predecir si a un individuo se le debe otorgar un préstamo, evaluando ciertos atributos para facilitar el proceso de selección de candidatos.

Se realizó un análisis comparativo entre dos algoritmos, Random Forest y Árboles de Decisión, utilizando el mismo conjunto de datos. Los resultados mostraron que el algoritmo Random Forest superó con una

presión del 80%, mientras que el algoritmo de Árboles de Decisión proporcionó una precisión del 73%. Este enfoque puede ayudar a las autoridades bancarias a tomar decisiones más informadas y a reducir el riesgo de incumplimiento de préstamos.

11. Un estudio sobre el modelado de puntuación de crédito con diferentes enfoques de selección de características y aprendizaje automático.

Según (Trivedi, 2020), un desafío significativo para las instituciones financieras es identificar a los candidatos adecuados para otorgar una línea de crédito sin asumir riesgos innecesarios. Para tomar decisiones tan cruciales, es vital analizar los datos demográficos y financieros previos de los solicitantes y construir un modelo automatizado de predicción de puntaje crediticio basado en inteligencia artificial y aprendizaje automático.

En este estudio, se utilizaron datos crediticios alemanes disponibles públicamente. Se demostró una mejora en la predicción de la calificación crediticia mediante el uso de diversas técnicas de selección de características, como ganancia de información, ratio de ganancia y chi-cuadrado, junto con clasificadores de aprendizaje automático, incluidos Bayesiano, Naïve Bayes, Random Forest, árbol de decisión (C5.0) y SVM (Máquina de Soporte Vectorial).

Se emplearon diferentes métricas de evaluación, como precisión, medida F, tasa de falsos positivos, tasa de falsos negativos y tiempo de entrenamiento, para analizar el rendimiento de los modelos. El estudio encontró que la combinación de Random Forest (RF) y Chi-Cuadrado (CS) ofrece un buen equilibrio, logrando una alta precisión, una excelente medida F y bajas tasas de falsos positivos y falsos negativos.

12. Predicción de riesgo crediticio en Colombia usando técnicas de inteligencia artificial

De acuerdo con (Borrero-Tigreros & Bedoya-Leiva, 2020), proponen modelos basados en tres técnicas de aprendizaje supervisado (redes neuronales, árboles de decisión y máquinas de soporte vectorial) para

predecir el próximo pago de la cuota de un cliente utilizando datos básicos de la operación, del cliente y de pagos de cuotas anteriores.

El criterio de comparación utilizado para los modelos predictivos de riesgo crediticio fue el análisis de la curva ROC, calculando el área bajo la curva (AUC) como indicador de la capacidad predictiva de cada modelo.

De acuerdo con los resultados obtenidos, los algoritmos Random Forest, C5.0, C4.5 y redes neuronales alcanzaron áreas bajo la curva de 88.29%, 80.56%, 78.24% y 69.91%, respectivamente. En contraste, el modelo basado en SVM no mostró el mismo rendimiento, con un AUC de 59.30%.

13. Evaluación del riesgo crediticio de fintech para pymes: evidencia de China.

Al analizar 1,8 millones de transacciones de préstamos de un banco en línea líder en China, (Huang et al., 2020) compara el enfoque fintech para evaluar el riesgo crediticio utilizando big data y modelos de aprendizaje automático con el enfoque bancario que utiliza datos financieros tradicionales y modelos de cuadros de mando.

Los datos utilizados para este análisis fueron proporcionados por MYbank de forma confidencial y no están disponibles al público dada la protección de la privacidad de los clientes.

El resultado del análisis y la sustitución de los modelos de cuadro de mando por modelos de aprendizaje automático como Random Forest mejoran significativamente la precisión del pronóstico (con el AUC aumentando de 0,74 a 0,87).

14. Integración de algoritmos de aprendizaje automático supervisados y no supervisados para la evaluación del riesgo crediticio.

Según (Bao et al., 2019), el problema de discriminar entre "buenos" y "malos" solicitantes de crédito en las instituciones financieras puede abordarse mediante una estrategia que combine el aprendizaje supervisado y no supervisado para la evaluación del riesgo crediticio. Su metodología incluye la obtención y limpieza de datos cuyo conjunto de datos a utilizar es créditos P2P chino, la selección de características óptimas, la implementación de un algoritmo de aprendizaje, el ajuste del

modelo y su evaluación. Los resultados confirmaron la superioridad de la integración propuesta de ambos tipos de algoritmos de aprendizaje automático.

15. Modelado predictivo para la detección de fraudes con tarjetas de crédito mediante análisis de datos.

Para (Patil et al., 2018), el sector financiero y bancario es fundamental en nuestra sociedad actual, donde casi todos interactúan con bancos, ya sea físicamente o en línea. Este estudio comparó tres modelos de aprendizaje automático: regresión logística, árbol de decisión y Random Forest, utilizando un conjunto de datos sobre fraude con tarjetas de crédito. La precisión de cada modelo se evaluó mediante una matriz de confusión, que muestra cómo se clasifican correctamente las tuplas en los modelos de entrenamiento y prueba.

Los modelos se evaluaron en función de parámetros como precisión, recuperación, exactitud y puntuación F1. El modelo Random Forest mostró un mejor rendimiento en comparación con la regresión logística y el árbol de decisión, destacándose en términos de exactitud, precisión y recuperación.

Tabla 1

Matriz de Investigaciones Similares

Cita	Problema abordado	Fuente de Datos Utilizado	Metodología	Resultados	Implicaciones
(Zhou et al., 2023)	Modelo de calificación crediticia de dos etapas basado en Random Forest	Datos de préstamos de 3045 pequeñas empresas en China	• Limpieza de datos • Utilizar SMOTE para tratar los datos desequilibrados • Construcción de características de crédito basada en Random Forest.	La metodología propuesta supera a otros modelos de aprendizaje automático.	
(Doko et al., 2021)	Evaluar diferentes modelos de aprendizaje automático para la evaluación del riesgo de crédito	Conjunto de datos del registro de crédito real del Banco Central de la República de Macedonia del Norte	• Limpieza de datos • Selección de características • Crear entrenamiento y conjunto de pruebas • Predicción • Evaluar el modelo	Los modelos que mejor puntuación F1 tienen, son: árbol de decisión, Random Forest y regresión lineal	Después dividir el conjunto de datos de entrenamiento y pruebas, se encontró que ambos conjuntos están muy desequilibrados.

(Freire López, 2021)	Modelo de clasificación de riesgo crediticio utilizando Random Forest en financiera del Ecuador	Conjunto de datos de Insotec	• Recopilación y preparación de datos • Selección de características • Creando el entrenamiento y conjunto de pruebas • Predicción • Evaluar el modelo	Los resultados de Random forest fueron los mejores para el conjunto de datos, se obtuvo una precisión del 97,2% y una tasa de error del 2.8%.
(Borrero-Tigreros & Bedoya-Leiva, 2020)	Identificar clientes que podrían incurrir en un estado de mora generando un posible riesgo de crédito para las entidades financieras	La base de datos es anónima, la cual cuenta con las tablas Clientes, Operaciones y DetIngresosClientes	• Selección de datos • Construcción de los modelos de predicción • Resultados	Los resultados obtenidos, de los algoritmos Random Forest y redes neuronales alcanzan áreas bajo la curva de 88.29%, y 69.91%, respectivamente

IDENTIFICACIÓN DEL OBJETO DE ESTUDIO

El fraude financiero es un problema significativo para las instituciones financieras no solamente para Ecuador sino a nivel mundial. Este fenómeno no solo genera pérdidas económicas sustanciales, sino que también afecta la confianza de los clientes y la reputación de las instituciones.

En el contexto actual, las instituciones financieras manejan un volumen enorme de transacciones diarias, tanto en línea como fuera de línea. Este alto volumen de datos, combinado con la necesidad de procesar transacciones en tiempo real, complica la identificación de patrones fraudulentos. Los métodos tradicionales de detección de fraude, basados en reglas predefinidas y análisis manual, no son lo suficientemente rápidos ni precisos para lidiar con el volumen y la complejidad de las transacciones modernas. Esto resulta en una alta tasa de falsos positivos y negativos, lo que significa que las transacciones legítimas pueden ser bloqueadas y las fraudulentas no detectadas.

PLANTEAMIENTO DEL PROBLEMA

El problema radica en la creciente sofisticación y frecuencia del fraude financiero, que afecta la capacidad de las instituciones financieras para detectar y prevenir actividades fraudulentas de manera efectiva. El uso de métodos tradicionales, como la revisión manual y las reglas predefinidas, no es suficiente para manejar el volumen y la complejidad de las transacciones modernas.

Este problema es crítico para la organización debido a varias razones:

- Pérdidas Financieras Significativas: Las transacciones fraudulentas no detectadas pueden resultar en pérdidas económicas considerables.
- Daño a la Reputación: La incapacidad de prevenir el fraude puede erosionar la confianza de los clientes y dañar la reputación de la organización.
- Costos Operativos Altos: Los métodos tradicionales son costosos y requieren una cantidad considerable de recursos humanos.
- Cumplimiento Normativo: Las instituciones deben cumplir con estrictas regulaciones para prevenir el fraude, y el incumplimiento puede llevar a sanciones severas.

Adoptar un enfoque analítico, especialmente mediante el uso de ML, está justificado por varias razones:

- Precisión y Eficiencia: Los modelos de ML pueden analizar grandes volúmenes de datos en tiempo real, identificando patrones anómalos con alta precisión.
- Automatización del Proceso: La automatización reduce la necesidad de intervención manual, optimizando los recursos y reduciendo costos operativos.
- Adaptabilidad: Los modelos pueden ajustarse y mejorar continuamente con nuevos datos, manteniéndose actualizados frente a nuevas tácticas de fraude.
- Mejora en la Toma de Decisiones: Proporciona información detallada y basada en datos, ayudando a la organización a tomar decisiones informadas y estratégicas.

Este enfoque no solo aborda eficazmente el problema del fraude financiero, sino que también fortalece la posición de la organización en términos de seguridad, cumplimiento y competitividad.

OBJETIVO GENERAL

Desarrollar un modelo de ML que pueda predecir transacciones fraudulentas en tiempo real para una empresa auxiliar del sistema financiero. Esto ayudará a la empresa a mitigar riesgos financieros y a proteger tanto sus activos como los de sus clientes.

OBJETIVOS ESPECÍFICOS

- Determinar un conjunto de datos relevante y representativo de las transacciones financieras.
- Evaluar y comparar al menos tres modelos de ML diferentes, utilizando métricas específicas de rendimiento como precisión, recall, F1-score y ROC-AUC para determinar su eficacia en la detección de fraudes.
- Analizar los resultados de las evaluaciones de los modelos y seleccionar aquel que demuestre la mayor precisión y eficacia en la predicción de transacciones fraudulentas.
- Crear una interfaz que permita la comunicación entre el modelo de ML y el sistema de gestión de crédito de la organización.

JUSTIFICACIÓN Y APLICACIÓN DE LA METODOLOGÍA

La predicción de fraude financiero es un desafío crítico para las instituciones financieras debido a las implicaciones económicas y de reputación que el fraude conlleva. La implementación de técnicas de ML para la detección de fraude es una estrategia prometedora que aprovecha el poder de los datos y los algoritmos avanzados para identificar patrones sospechosos y prevenir actividades fraudulentas.

En el presente proyecto capstone, se justifica la utilización de una metodología basada en recolección de datos, limpieza, pre-procesamiento y/o transformación de datos, identificación y descripción de variables, visualización de variables y Selección de modelo estadístico (Sahoo et al., 2019), como se detalla a continuación.

1. Recolección de datos

La primera etapa en cualquier proyecto de análisis de datos es la recolección de los datos. Para un proyecto de predicción de fraude financiero, los datos pueden ser recolectados de diversas fuentes, como bases de datos de transacciones de la institución financiera, registros históricos de fraudes, y datos externos de comportamiento financiero. Sin embargo, para el presente proyecto el conjunto de datos seleccionado proviene de Kaggle.

Kaggle es una plataforma que reúne una comunidad de ciencia de datos, en la cual se encuentran disponibles miles de conjuntos de datos de todas las áreas, los cuales sirven para estudios con diferentes fines estudiantiles o profesionales. El conjunto de datos seleccionado se llama “Datos de transacciones fraudulentas” (*Fraudulent Transactions Data*, n.d.), la cual se encuentra en un formato csv. Se seleccionó este conjunto de datos debido a la gran cantidad de información que proporciona sobre diferentes transacciones a lo largo del tiempo, lo cual es sumamente útil para estudios de predicción financiera.

Análisis Exploratorio de Datos (EDA)

1. Información del Dataset:
    - El dataset contiene 6,362,620 registros y 11 columnas.
    - No hay valores nulos en el dataset.
    - Las variables incluyen tanto variables numéricas como categóricas.
2. Estadísticas Descriptivas:
    - La media del monto (amount) de las transacciones es de aproximadamente 179,472.
    - Las variables de saldo antiguo y nuevo (saldo_antes_origen, saldo_despues_origen, saldo_antes_destino, saldo_despues_destino) tienen una gran variabilidad, lo que sugiere la presencia de transacciones de diversos tamaños.
    - La variable es_fraude tiene una distribución altamente desbalanceada, con solo el 0.129% de las transacciones marcadas como fraudulentas.
3. Distribución de la Variable Objetivo (es_fraude):
    - El 99.871% de las transacciones no son fraudulentas, mientras que solo el 0.129% son fraudulentas.

2. Limpieza, pre-procesamiento y/o transformación de datos

La calidad de los datos es fundamental para el rendimiento de los modelos de ML. La limpieza de datos implica la eliminación de valores nulos, la corrección de errores y la transformación de variables categóricas. En este estudio, se aplicará one-hot encoding para las variables categóricas, lo que permite que los algoritmos de ML procesen adecuadamente las características categóricas (Yu et al., 2022).

3. Identificación y descripción de variables

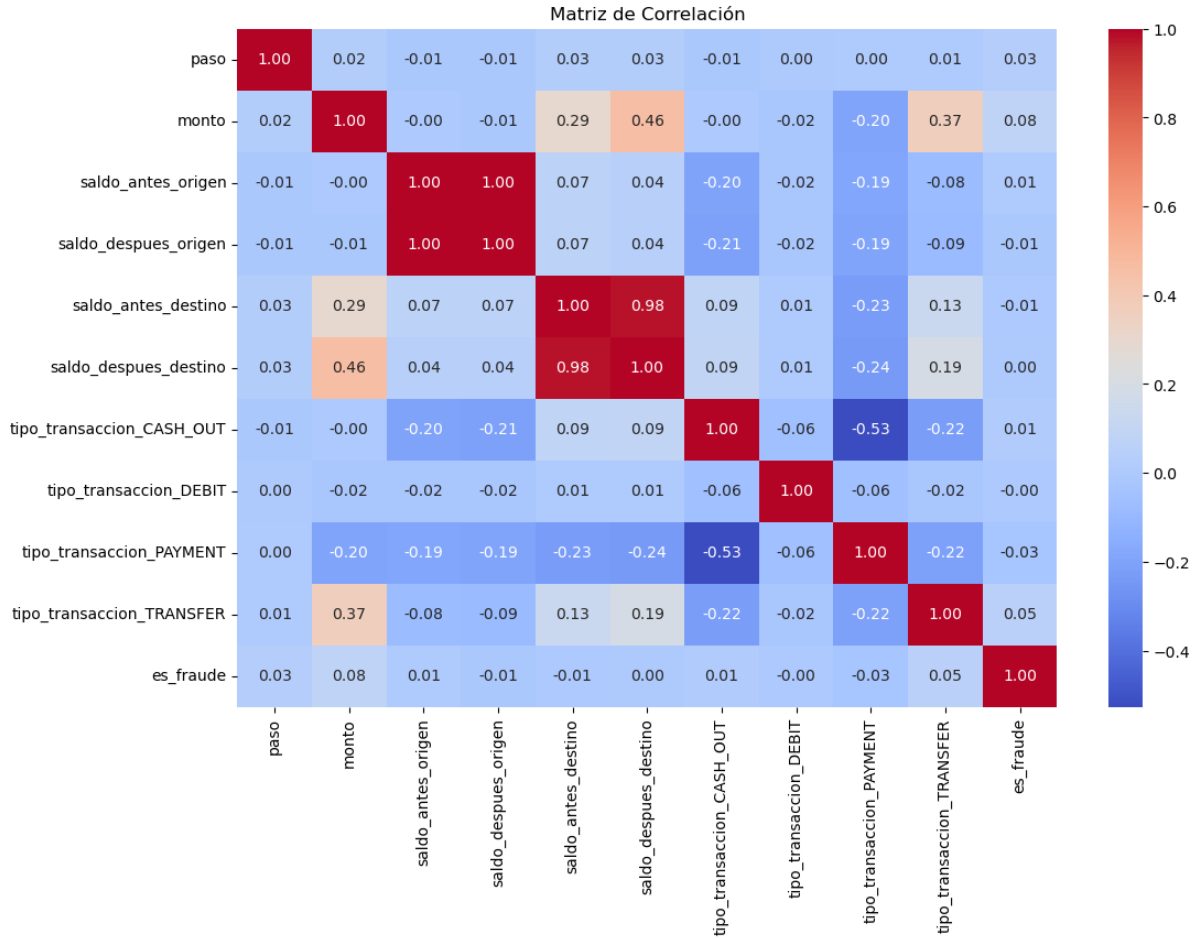
La selección de variables es un proceso para identificar las características más relevantes que influyen en la predicción del fraude.

A continuación, se describe las variables del conjunto de datos seleccionado y su matriz de correlación, con el fin de poder identificar la intensidad y la dirección de la relación entre las variables.

Tabla 2
Matriz de Descripción de Variables

Variable	Fuente de datos	Tipo de dato	Descripción
paso	Kaggle	Numérica (entero)	Número del paso de tiempo en la simulación de transacciones.
tipo_transaccion	Kaggle	Categórica	Tipo de transacción (e.g., PAYMENT, TRANSFER, CASH_OUT, etc.).
monto	Kaggle	Numérica (flotante	Monto de la transacción.
nombre_cliente_origen	Kaggle	Categórica	Identificador del cliente que origina la transacción.
saldo_antes_origen	Kaggle	Numérica (flotante)	Saldo del cliente antes de la transacción.
saldo_despues_origen	Kaggle	Numérica (flotante)	Saldo del cliente después de la transacción.
nombre_cliente_destino	Kaggle	Categórica	Identificador del cliente que recibe la transacción.
saldo_antes_destino	Kaggle	Numérica (flotante)	Saldo del destinatario antes de la transacción.
saldo_despues_destino	Kaggle	Numérica (flotante)	Saldo del destinatario después de la transacción.
es_fraude	Kaggle	Binaria	Indicador de si la transacción es fraudulenta (1) o no (0).
es_marcado_fraude	Kaggle	Binaria	Indicador de si la transacción fue marcada como potencial fraude (1) o no (0).

Figura 1
Matriz de Correlación



Naturaleza de los Datos

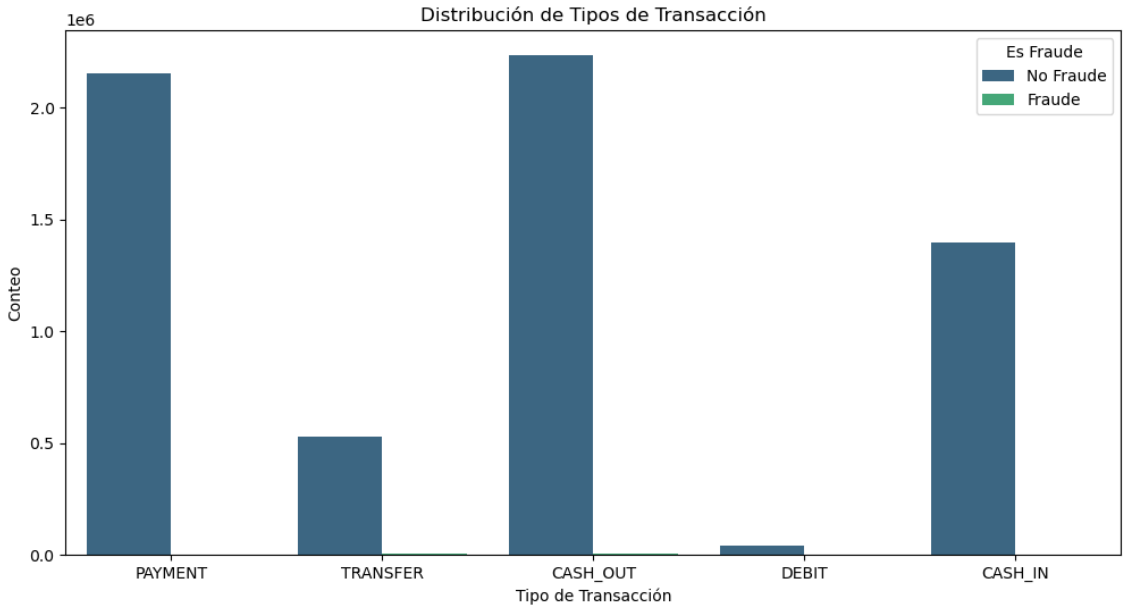
- Temporalidad: Los datos recolectados representan transacciones realizadas en diferentes momentos en el tiempo, indicadas por la variable “paso”. Esto es importante para detectar patrones temporales en el fraude.
- Transaccionalidad: Las variables como monto, saldo_antes_origen, saldo_despues_origen, saldo_antes_destino y saldo_despues_destino son datos transaccionales que describen las características financieras de cada transacción.
- Identificadores de Cliente: nombre_cliente_origen y nombre_cliente_destino son identificadores únicos para los clientes que participan en la transacción. Estos datos pueden ser usados para realizar análisis de comportamiento del cliente.
- Indicadores de Fraude: es_fraude y es_marcado_fraude son variables binarias que indican si una transacción es fraudulenta o si ha sido marcada como sospechosa por el sistema.

4. Visualización de variables

La visualización de los datos es un paso crucial para entender las distribuciones y las relaciones entre las variables. Técnicas como los gráficos de caja (boxplots), gráficos de dispersión e histogramas pueden revelar patrones, tendencias y valores atípicos que pueden ser indicativos de fraude. La visualización también facilita la comunicación de los hallazgos a las partes interesadas.

Figura 2

Distribución por Tipo de Transacción



5. Selección del modelo estadístico

Regresión Logística

La regresión logística es adecuada para problemas de clasificación binaria como la detección de fraude. Es fácil de interpretar y proporciona probabilidades de pertenencia a cada clase.

Justificación:

- Interpretabilidad: Proporciona coeficientes que pueden interpretarse en términos de probabilidades.
- Eficacia: Bien documentada y efectiva para problemas de clasificación binaria.

Figura 3
Ecuación del Modelo de Regresión Logística

$$log(1 - P(Fraude)P(Fraude)) = \beta 0 + \beta 1 \cdot paso + \beta 2 \cdot monto + \beta 3 \cdot saldoAntiguoOrg + \beta 4 \cdot nuevoSaldoOrg + \beta 5 \cdot saldoAntiguoDest + \beta 6 \cdot nuevoSaldoDest + \beta 7 \cdot tipo\_CASH\_OUT + \beta 8 \cdot tipo\_DEBIT + \beta 9 \cdot tipo\_PAYMENT + \beta 10 \cdot tipo\_TRANSFER$$

Árbol de Decisión

Los árboles de decisión son modelos intuitivos y fáciles de interpretar. Pueden manejar relaciones no lineales y capturar interacciones entre variables.

Justificación:

- Interpretabilidad: Los árboles de decisión son fácilmente interpretables visualmente.
- Manejo de No Linealidades: Captura relaciones no lineales entre variables.

Ecuación del Árbol de Decisión:

No tiene una forma matemática explícita, pero se basa en dividir el espacio de características en regiones homogéneas respecto a la variable objetivo.

Random Forest

El Random Forest es un algoritmo de ensamble que utiliza múltiples árboles de decisión. Es robusto y maneja bien datos desbalanceados.

Justificación:

- Robustez: Menos propenso al sobreajuste comparado con un solo árbol de decisión.
- Eficacia en Datos Desbalanceados: Puede manejar el desbalance de clases mediante ponderación de clases.

Ecuación del Modelo de Random Forest:

Al igual que los árboles de decisión, no tiene una forma matemática explícita. La predicción final se hace por agregación (promedio o votación) de las predicciones individuales de los árboles.

Justificación sobre la selección de Variables

La selección de variables se basó en la literatura revisada y en el análisis exploratorio de datos. Las variables seleccionadas tienen relevancia teórica y empírica en la detección de fraudes financieros:

- Paso: Indica el momento en el tiempo de la transacción.
- Monto: El monto de la transacción es crucial para identificar anomalías.
- Saldo Antiguo y Nuevo (Origen y Destino): Permiten analizar el comportamiento del saldo antes y después de la transacción.
- Tipo de Transacción: Diferentes tipos de transacciones pueden tener diferentes perfiles de riesgo.

La selección de los modelos de Regresión Logística, Árbol de Decisión y Random Forest está justificada por su eficacia comprobada en la literatura y su capacidad para manejar problemas de clasificación binaria y desbalance de clases. La selección de variables se basa en su relevancia teórica y empírica en la detección de fraudes financieros, asegurando que los modelos construidos sean robustos y efectivos para predecir transacciones fraudulentas.

RESULTADOS

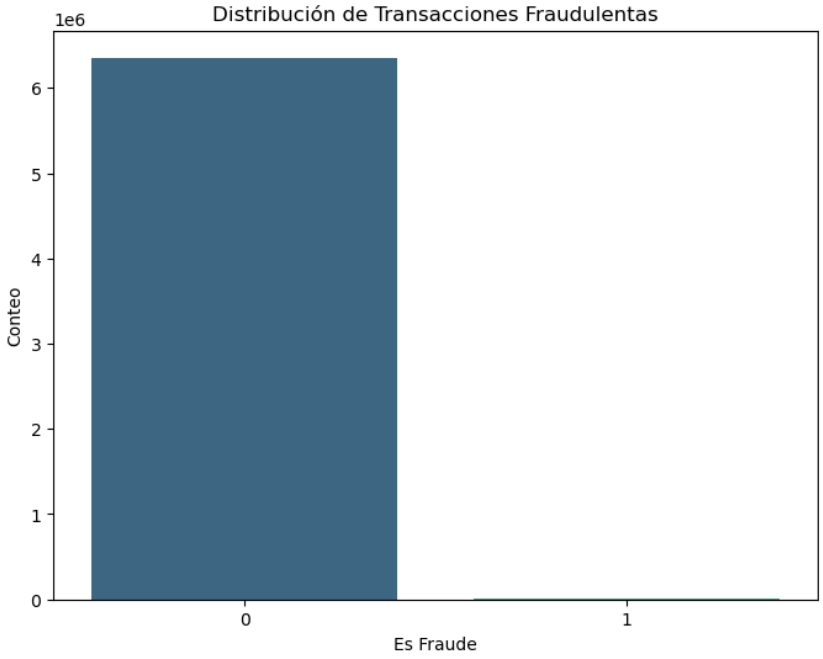
1. Análisis del modelo estadístico

De acuerdo con el análisis exploratorio de datos en el software de licencia libre como es Python, se cuenta con las siguientes visualizaciones que ayudan a entender de mejor manera la información:

- Distribución de las Transacciones Fraudulentas:
  - La mayoría de las transacciones no son fraudulentas, lo cual indica un desbalance significativo en el conjunto de datos.
  - Estos resultados demuestran que se trata de datos muy desequilibrados, ya que el número de transacciones fraudulentas es 8213 y el número de transacciones legítimas = 6354407. Por lo tanto, Random Forest, Decision Trees, Redes neuronales y XGBoost son buenos métodos para los datos altamente desequilibrados. (Moreno et al., 2023)

Figura 4

Distribución de Transacciones Fraudulentas



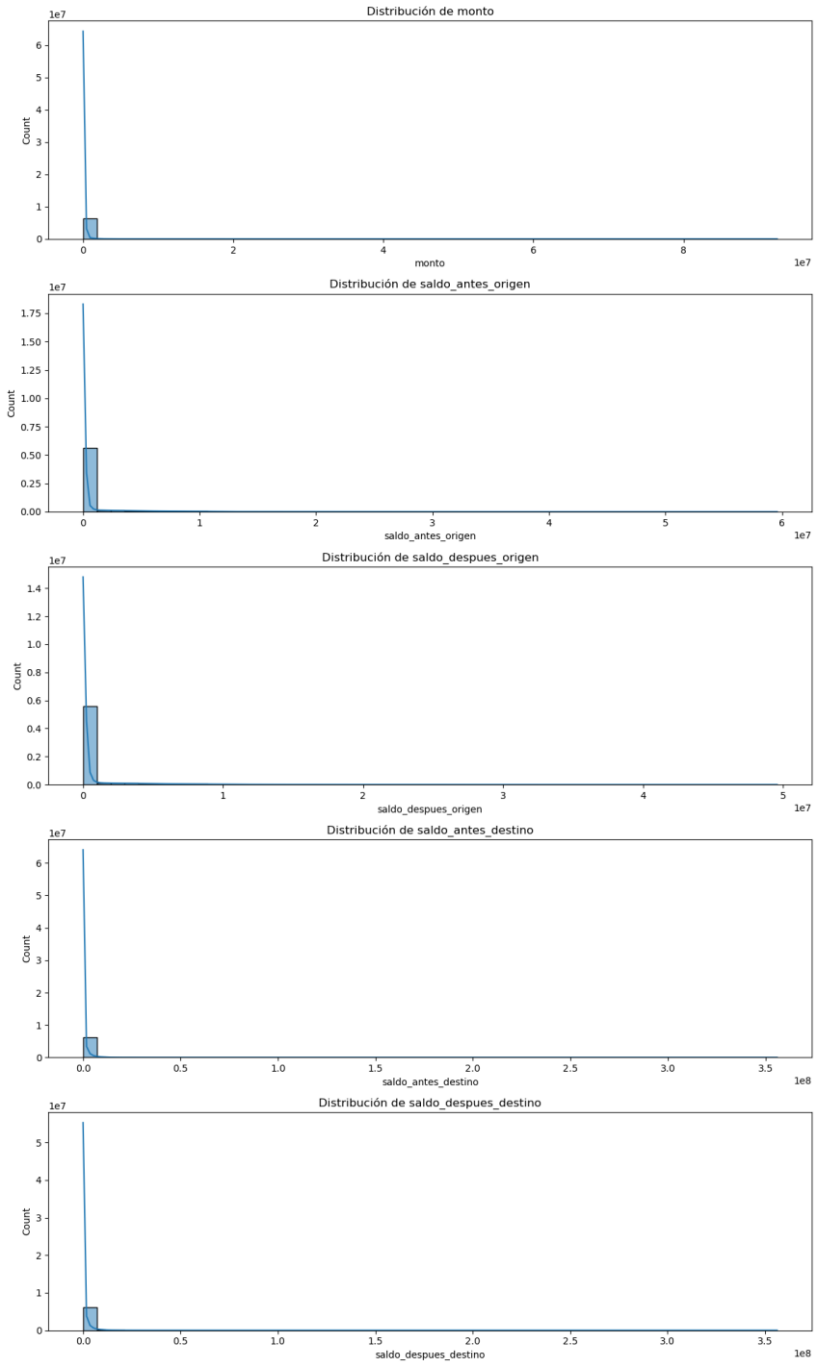
El gráfico de barras ilustra la distribución de transacciones, distinguiendo entre transacciones no fraudulentas (etiquetadas como "0") y fraudulentas (etiquetadas como "1"). La barra para las transacciones no fraudulentas es

significativamente más alta, con un conteo que supera los seis millones, mientras que la barra para las transacciones fraudulentas es casi imperceptible, indicando una cantidad extremadamente baja en comparación. Este gráfico destaca la marcada disparidad entre la cantidad de transacciones legítimas y fraudulentas, mostrando que las transacciones fraudulentas representan una fracción muy pequeña del total.

- Distribución de las Variables Numéricas:
  - Las distribuciones de monto, saldo_antes_origen, saldo_despues_origen, saldo_antes_destino, y saldo_despues_destino muestran que hay una gran variabilidad en los datos transaccionales.
  - Los histogramas indican la presencia de valores extremos en varias de estas variables, lo cual es común en datos financieros.

Figura 5

Distribución de Variables Numéricas

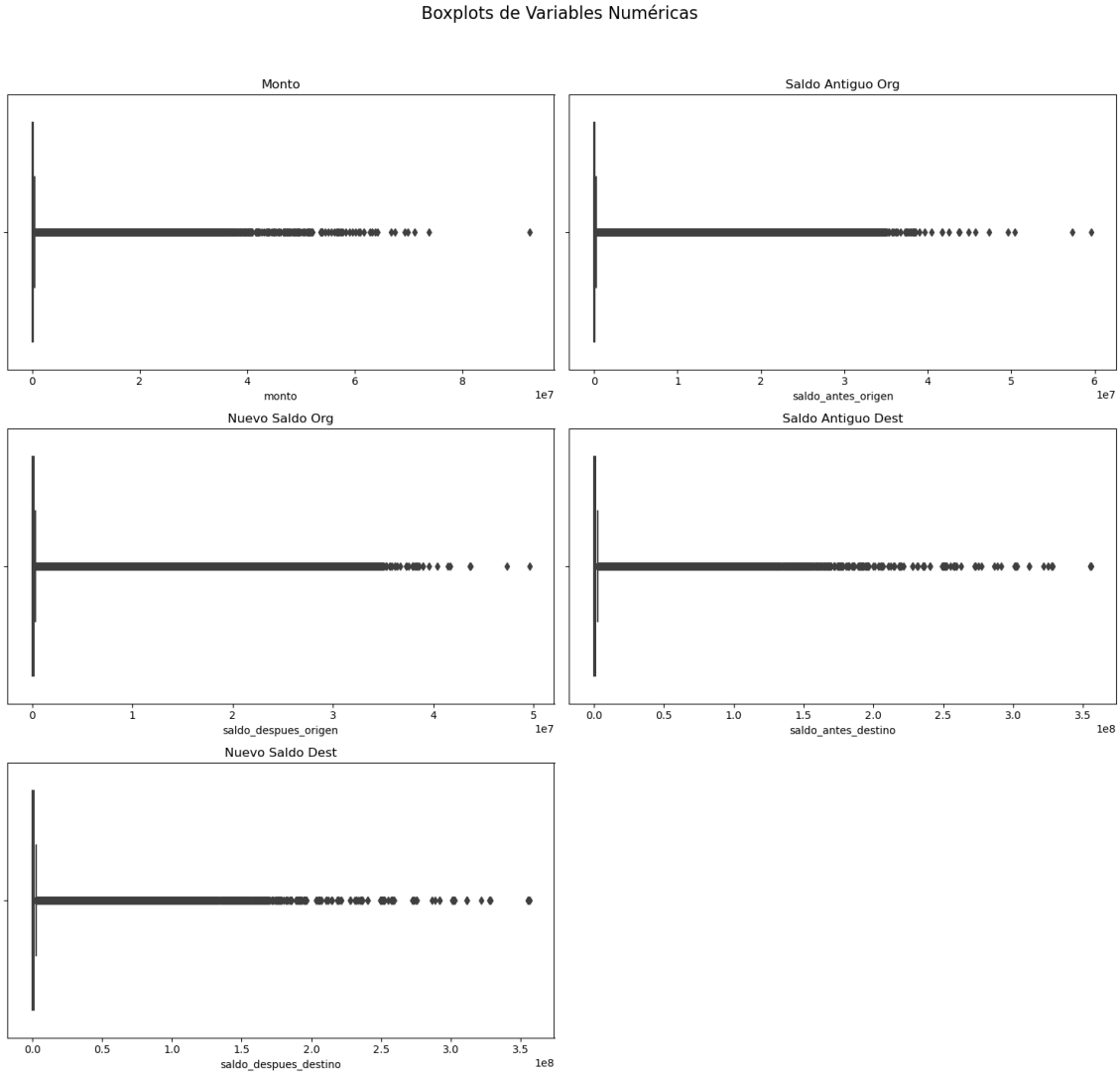


La imagen presenta histogramas para las variables: monto, saldo_antes_origen, saldo_despues_origen, saldo_antes_destino, y saldo_despues_destino. En todos los casos, la mayoría de los datos se concentran en valores bajos, con una larga cola hacia la derecha, indicando la presencia de unos pocos valores extremadamente altos. Esta distribución sesgada sugiere que, aunque la

mayoría de los montos y saldos son relativamente pequeños, existen casos de montos y saldos muy grandes que afectan la distribución general de los datos.

- Valores Atípicos:
    - Los gráficos de caja (boxplots) revelan la presencia de valores atípicos en las variables numéricas. Esto es relevante ya que los valores atípicos pueden ser indicativos de comportamiento fraudulento.
    - Variables como monto y saldo_antes_origen muestran una gran cantidad de valores atípicos, que deben ser considerados en el modelado.

Figura 6
Valores Atípicos



La figura muestra cinco gráficos de caja (boxplots) para las variables numéricas: Monto, Saldo Antiguo Org, Nuevo Saldo Org, Saldo Antiguo Dest, y Nuevo Saldo Dest. Cada gráfico revela que la mayoría de los datos se concentran en valores bajos, con numerosos valores atípicos en el extremo derecho. Esto indica una distribución sesgada con algunos valores extremadamente altos en todas las variables. Los gráficos evidencian la presencia de transacciones o saldos muy elevados en un pequeño número de casos, mientras que la mayoría de los datos se mantienen en rangos más bajos.

El análisis exploratorio de datos proporciona una comprensión profunda de las características de las transacciones financieras y las variables relacionadas con el fraude. Las visualizaciones revelan patrones importantes y valores atípicos que son fundamentales para la construcción de modelos predictivos eficaces. Este análisis establece una base sólida para la siguiente fase del proyecto, que es Selección del modelo estadístico para el fraude financiero utilizando técnicas de ML.

2. Interpretación de resultados

Los modelos de Regresión Logística, Árbol de Decisión y Random Forest proporcionan una comprensión detallada de las relaciones entre las variables y su impacto en la probabilidad de fraude financiero. La Regresión Logística ofrece coeficientes y significancia estadística clara, los Árboles de Decisión brindan una visualización intuitiva y las Importancias de Variables en Random Forest identifican las características más influyentes.

Estas técnicas de análisis y sus interpretaciones permiten no solo predecir fraudes de manera efectiva, sino también comprender los factores subyacentes que contribuyen al fraude, lo cual es crucial para diseñar estrategias preventivas y mejorar los sistemas de detección de fraude en instituciones financieras.

A continuación, se evalúa los modelos en base al reporte de clasificación que proporciona una serie de métricas para evaluar el rendimiento de los modelos en la tarea de detección de fraudes financieros.

1. Regresión Logística

Con base en el resultado de este modelo, se realiza el siguiente análisis:

Tabla 3
Reporte Regresión Logística

	Precisión	Recall	F1-Score	Support
0	1.00	0.96	0.98	1906322
1	0.03	0.92	0.06	2464
ROC-AUC	0.98521			

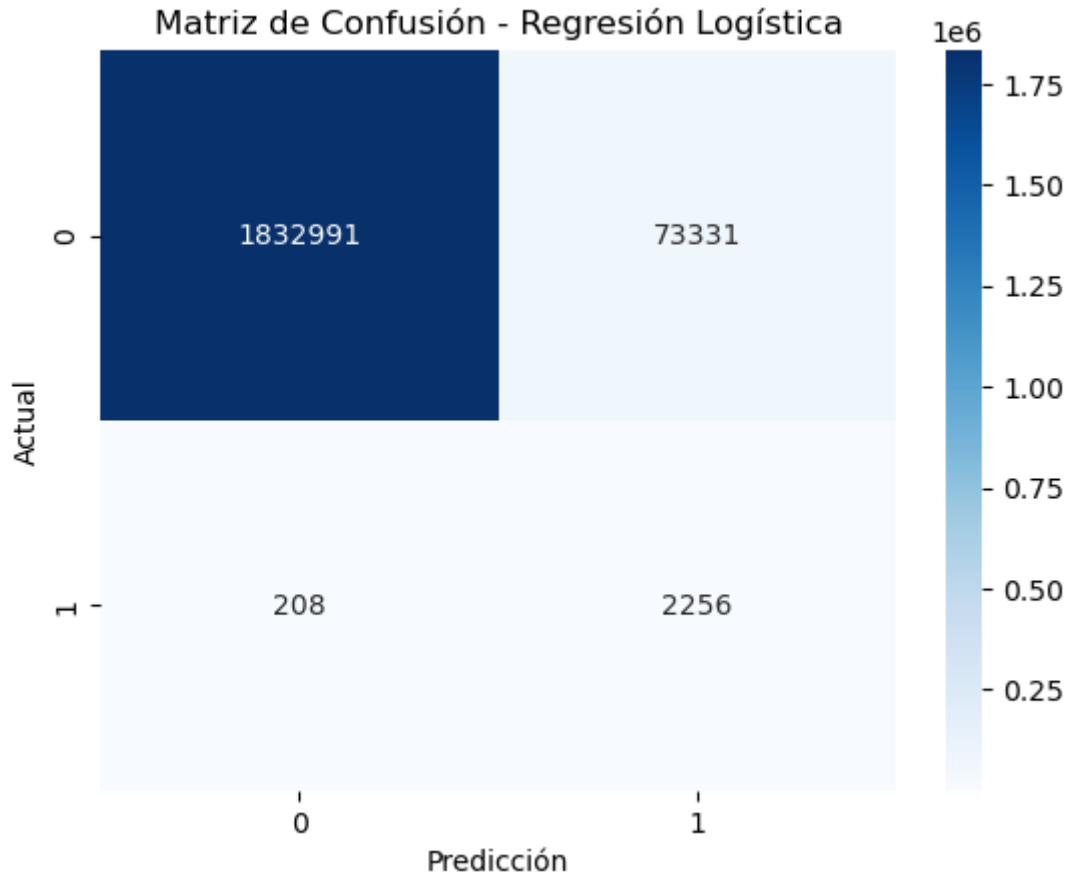
1. Precision (Precisión):
    - Clase 0 (No Fraude): 1.00

- Clase 1 (Fraude): 0.03
- Interpretación: La precisión es perfecta para las transacciones no fraudulentas (1.00), lo que significa que todas las transacciones predichas como no fraudulentas realmente lo son. Sin embargo, la precisión para las transacciones fraudulentas es muy baja (0.03), lo que indica que la mayoría de las transacciones clasificadas como fraudulentas no lo son en realidad.

2. Recall (Sensibilidad o Exhaustividad):
   - Clase 0 (No Fraude): 0.96
   - Clase 1 (Fraude): 0.92
   - Interpretación: El recall para las transacciones no fraudulentas es alto (0.96), lo que indica que el modelo identifica correctamente la mayoría de las transacciones no fraudulentas. El recall para las transacciones fraudulentas también es alto (0.92), lo que sugiere que el modelo es capaz de identificar la mayoría de las transacciones fraudulentas.
3. F1-Score:
   - Clase 0 (No Fraude): 0.98
   - Clase 1 (Fraude): 0.06
   - Interpretación: El F1-Score es una medida combinada de precisión y recall. Para la clase no fraudulenta, es muy alto (0.98), lo que indica un buen balance entre precisión y recall. Sin embargo, para la clase fraudulenta, el F1-Score es muy bajo (0.06), debido a la baja precisión.
4. Support (Soporte):
   - Clase 0 (No Fraude): 1,906,322
   - Clase 1 (Fraude): 2,464
   - Interpretación: El soporte indica el número de ocurrencias de cada clase en el conjunto de datos. Hay una gran desproporción entre las clases, con muchas más transacciones no fraudulentas que fraudulentas.

A continuación, se presenta el gráfico de la matriz de confusión:

Figura 7
Matriz de Confusión - Regresión Logística



Con base en esta matriz, se presenta el siguiente análisis de resultados:

- True Negatives (TN): 1,832,991
  Número de transacciones no fraudulentas correctamente clasificadas como no fraudulentas.
- False Positives (FP): 73,331
  Número de transacciones no fraudulentas incorrectamente clasificadas como fraudulentas.
- False Negatives (FN): 208
  Número de transacciones fraudulentas incorrectamente clasificadas como no fraudulentas.
- True Positives (TP): 2,256
  Número de transacciones fraudulentas correctamente clasificadas como fraudulentas.

2. Árbol de Decisión

Con base en el resultado de este modelo, se realiza el siguiente análisis:

Tabla 4
Reporte Árbol de Decisión

	Precisión	Recall	F1-Score	Support
0	1.00	1.00	1.00	1906322
1	0.87	0.85	0.86	2464
ROC-AUC	0.92260			

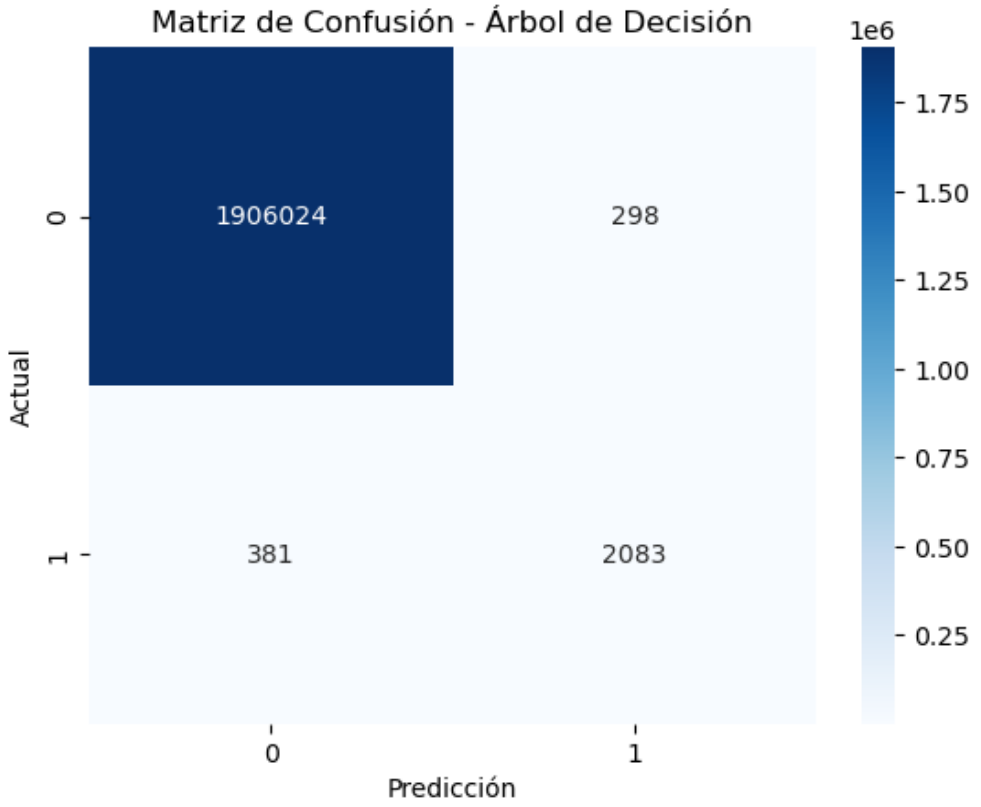
1. Precision (Precisión):
   - Clase 0 (No Fraude): 1.00
   - Clase 1 (Fraude): 0.87
   - Interpretación: La precisión es perfecta para las transacciones no fraudulentas (1.00), lo que significa que todas las transacciones predichas como no fraudulentas realmente lo son. La precisión para las transacciones fraudulentas es alta (0.87), lo que indica que una gran proporción de las transacciones clasificadas como fraudulentas realmente lo son.
2. Recall (Sensibilidad o Exhaustividad):
   - Clase 0 (No Fraude): 1.00
   - Clase 1 (Fraude): 0.85
   - Interpretación: El recall es perfecto para las transacciones no fraudulentas (1.00), lo que indica que el modelo identifica correctamente todas las transacciones no fraudulentas. El recall para las transacciones fraudulentas es también alto (0.85), sugiriendo que el modelo es capaz de identificar la mayoría de las transacciones fraudulentas.
3. F1-Score:
   - Clase 0 (No Fraude): 1.00
   - Clase 1 (Fraude): 0.86
   - Interpretación: El F1-Score es perfecto para la clase no fraudulenta (1.00), lo que refleja un excelente balance entre precisión y recall. Para

la clase fraudulenta, el F1-Score es alto (0.86), lo que indica un buen equilibrio entre precisión y recall.

4. Support (Soporte):
   - Clase 0 (No Fraude): 1,906,322
   - Clase 1 (Fraude): 2,464
   - Interpretación: El soporte indica el número de ocurrencias de cada clase en el conjunto de datos. Hay una gran desproporción entre las clases, con muchas más transacciones no fraudulentas que fraudulentas.

A continuación, se presenta el gráfico de la matriz de confusión:

Figura 8
Matriz de Confusión - Árbol de Decisión



Con base en esta matriz, se presenta el siguiente análisis de resultados:

- True Negatives (TN): 1,906,024
  Número de transacciones no fraudulentas correctamente clasificadas como no fraudulentas.
- False Positives (FP): 298

Número de transacciones no fraudulentas incorrectamente clasificadas como fraudulentas.

- False Negatives (FN): 381
  Número de transacciones fraudulentas incorrectamente clasificadas como no fraudulentas.
- True Positives (TP): 2,083
  Número de transacciones fraudulentas correctamente clasificadas como fraudulentas.

3. Random Forest

Con base en el resultado de este modelo, se realiza el siguiente análisis:

Tabla 5
Reporte Random Forest

	Precisión	Recall	F1-Score	Support
0	1.00	1.00	1.00	1906322
1	0.98	0.77	0.86	2464
ROC-AUC	0.99480			

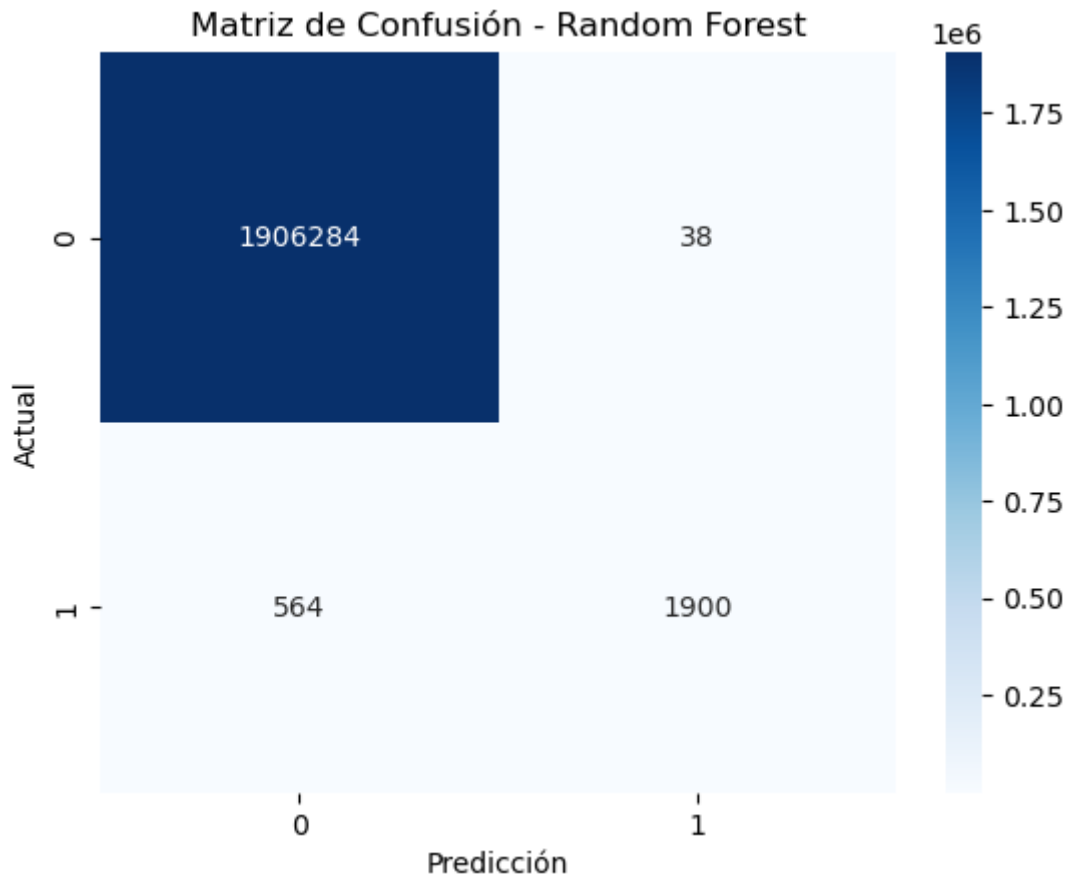
1. Precision (Precisión):
   - Clase 0 (No Fraude): 1.00
   - Clase 1 (Fraude): 0.98
   - Interpretación: La precisión es perfecta para las transacciones no fraudulentas (1.00), lo que significa que todas las transacciones predichas como no fraudulentas realmente lo son. La precisión para las transacciones fraudulentas es muy alta (0.98), lo que indica que casi todas las transacciones clasificadas como fraudulentas realmente lo son.
2. Recall (Sensibilidad o Exhaustividad):
   - Clase 0 (No Fraude): 1.00
   - Clase 1 (Fraude): 0.77
   - Interpretación: El recall es perfecto para las transacciones no fraudulentas (1.00), lo que indica que el modelo identifica correctamente todas las transacciones no fraudulentas. El recall para

las transacciones fraudulentas es alto (0.77), sugiriendo que el modelo es capaz de identificar la mayoría de las transacciones fraudulentas, aunque hay un margen de mejora.

3. F1-Score:
    - Clase 0 (No Fraude): 1.00
    - Clase 1 (Fraude): 0.86
    - Interpretación: El F1-Score es perfecto para la clase no fraudulenta (1.00), lo que refleja un excelente balance entre precisión y recall. Para la clase fraudulenta, el F1-Score es alto (0.86), lo que indica un buen equilibrio entre precisión y recall.
4. Support (Soporte):
    - Clase 0 (No Fraude): 1,906,322
    - Clase 1 (Fraude): 2,464
    - Interpretación: El soporte indica el número de ocurrencias de cada clase en el conjunto de datos. Hay una gran desproporción entre las clases, con muchas más transacciones no fraudulentas que fraudulentas.

A continuación, se presenta el gráfico de la matriz de confusión:

Figura 9
Matriz de Confusión - Random Forest

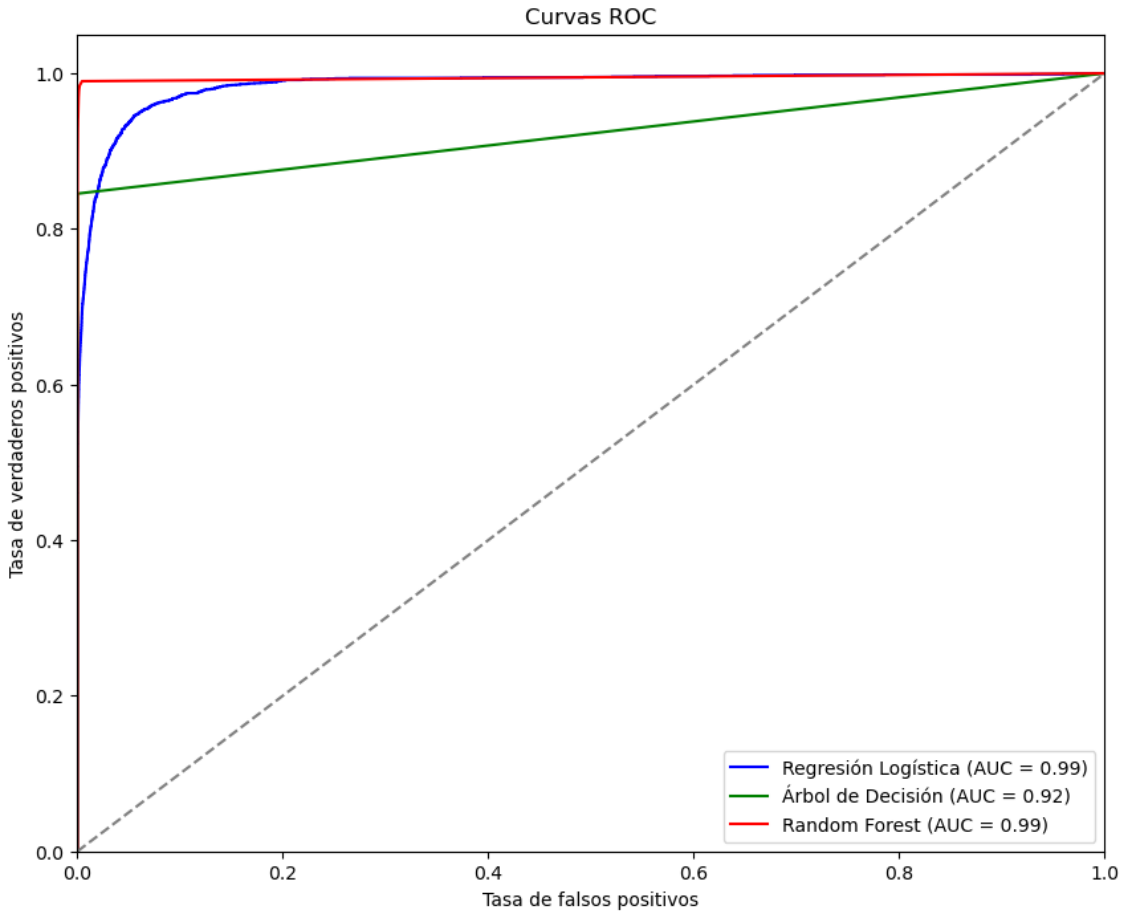


Con base en esta matriz, se presenta el siguiente análisis de resultados:

- True Negatives (TN): 1,906,284
  Número de transacciones no fraudulentas correctamente clasificadas como no fraudulentas.
- False Positives (FP): 38
  Número de transacciones no fraudulentas incorrectamente clasificadas como fraudulentas.
- False Negatives (FN): 564
  Número de transacciones fraudulentas incorrectamente clasificadas como no fraudulentas.
- True Positives (TP): 1,900
  Número de transacciones fraudulentas correctamente clasificadas como fraudulentas.

A continuación, se analiza una gráfica de curvas ROC, cuyo objetivo es determinar cuál de los tres modelos utilizados se ajusta mejor al conjunto de datos empleado para predecir el fraude financiero.

Figura 10
Curvas ROC



Interpretación de la gráfica:

La gráfica de curvas ROC (Receiver Operating Characteristic) es una herramienta visual que evalúa el rendimiento de los modelos de clasificación, especialmente en problemas de desbalance de clases como la detección de fraude. Cada curva ROC representa la relación entre la tasa de verdaderos positivos (True Positive Rate, TPR) y la tasa de falsos positivos (False Positive Rate, FPR) en varios umbrales de decisión.

1. Regresión Logística (AUC = 0.99):

- Curva (línea azul): La curva ROC del modelo de Regresión Logística también está cerca de la esquina superior izquierda, indicando un rendimiento muy bueno.
- AUC (Área bajo la curva): Un AUC de 0.99, similar al de Random Forest, sugiere que este modelo también tiene una alta capacidad de discriminación entre las clases.

2. Árboles de Decisión (AUC = 0.92):
    - Curva (línea verde): La curva ROC del modelo de Árbol de Decisión no es tan cercana a la esquina superior izquierda comparada con las otras dos, pero aún muestra un buen rendimiento.
    - AUC (Área bajo la curva): Un AUC de 0.92 es indicativo de un buen rendimiento, aunque no tan alto como los otros dos modelos.
3. Random Forest (AUC = 0.99):
    - Curva (línea roja): La curva ROC del modelo Random Forest está muy cerca de la esquina superior izquierda, indicando un rendimiento excelente.
    - AUC (Área bajo la curva): Un AUC de 0.99 indica que el modelo tiene una alta capacidad para distinguir entre clases fraudulentas y no fraudulentas. Cuanto más cerca esté el AUC de 1, mejor será el rendimiento del modelo.

De acuerdo con estos análisis, se determina los siguiente:

- Mejor Modelo: Tanto el Random Forest como la regresión logística tienen un AUC de 0.99, indicando que ambos modelos tienen un rendimiento excelente en términos de capacidad discriminativa. Sin embargo, el Random Forest tiene una ligera ventaja debido a su capacidad para manejar relaciones no lineales y su robustez en la clasificación.
- Árbol de Decisión: Aunque el árbol de decisión tiene un AUC de 0.92, que sigue siendo bueno, su rendimiento es inferior al de los otros dos modelos.
- Recomendación: Para la detección de fraudes financieros, tanto la regresión logística como el Random Forest son modelos adecuados debido a su alta capacidad discriminativa. Sin embargo, el Random Forest puede ser preferido debido a su robustez y flexibilidad.

De acuerdo con los análisis previos sobre los tres modelos y gracias a la gráfica ROC y los valores de AUC, se demuestra que el modelo Random Forest y la Regresión Logística son modelos altamente efectivos para la detección de fraudes, con una ligera preferencia por el Random Forest debido a sus características inherentes.

DISCUSIÓN DE LOS RESULTADOS Y PROPUESTA DE SOLUCIÓN

Contribución del Análisis de Datos y Random Forest

1. Detección Eficiente de Fraude:
    - Precisión y Recall Altos: El modelo Random Forest implementado ha mostrado una alta precisión (0.98) y un buen recall (0.77) para la clase fraudulenta, lo que indica su capacidad para identificar la mayoría de las transacciones fraudulentas con una baja tasa de falsos positivos. Estudios recientes han demostrado la eficacia de Random Forest en la detección de fraudes debido a su capacidad para manejar grandes conjuntos de datos y detectar patrones complejos (Li et al., 2020).
    - Capacidad Discriminativa: El ROC-AUC score de 0.9949 muestra que el modelo tiene una excelente capacidad para distinguir entre transacciones fraudulentas y no fraudulentas, lo que es crucial para minimizar tanto las pérdidas por fraude como las interrupciones para los clientes legítimos (Verster & Fourie, 2023).
2. Interpretabilidad y Robustez:
    - Importancia de Características: Random Forest proporciona información sobre la importancia de las variables, ayudando a entender qué características son más indicativas de fraude. Esto permite a la organización focalizarse en variables clave para la supervisión y control.
    - Robustez del Modelo: La técnica de Random Forest es robusta frente a datos ruidosos y variables no lineales, asegurando un rendimiento consistente incluso en condiciones cambiantes (Dávila-Morán et al., 2023).

Estrategia Organizacional Basada en Resultados

A partir de los resultados obtenidos con el modelo Random Forest, se puede diseñar una estrategia organizacional que no solo optimiza la detección de fraude, sino que también proporciona un marco para la toma de decisiones gerenciales.

1. Implementación del Sistema de Detección de Fraude
    - Integrar el modelo Random Forest en el sistema de gestión de transacciones en tiempo real.
    - Detectar y marcar transacciones sospechosas automáticamente, permitiendo una intervención rápida antes de completar la transacción.
2. Monitoreo Continuo y Actualización del Modelo
    - Establecer un ciclo de monitoreo y actualización continua del modelo.
    - Adaptar el modelo a nuevas técnicas de fraude mediante el reentrenamiento periódico con datos recientes, asegurando que el sistema se mantenga efectivo frente a nuevas amenazas (Nami & Shajari, 2018).
3. Formación y Capacitación del Personal
    - Capacitar al personal en el uso e interpretación de los resultados del modelo.
    - Aumentar la capacidad del equipo para identificar patrones de fraude y tomar decisiones informadas basadas en las predicciones del modelo.
4. Mejora de la Experiencia del Cliente
    - Minimizar las falsas alarmas ajustando los umbrales de decisión y refinando el modelo.
    - Reducir las interrupciones para los clientes legítimos, mejorando la satisfacción y reteniendo la confianza de los clientes (Li et al., 2020).
5. Comunicación y Colaboración Interdepartamental

- Facilitar la comunicación entre los departamentos de TI, análisis de datos y operaciones para implementar mejoras basadas en los hallazgos del modelo.
- Fomentar una cultura de datos en la organización donde las decisiones se basen en análisis y evidencias (Verster & Fourie, 2023).

6. Informe y Auditoría Regulares
    - Generar informes regulares sobre el rendimiento del modelo y los casos de fraude detectados.
    - Proporcionar a la gerencia y a los entes reguladores, la información clara sobre la eficacia del sistema de detección de fraude y cumplir con los requisitos de auditoría.

Implicaciones para la Toma de Decisiones Gerenciales

Asignación de Recursos: Basado en la precisión y recall del modelo, la gerencia puede justificar la inversión en tecnologías de ML para la detección de fraude, asignando recursos adicionales para mejorar y mantener el sistema.

- Estrategias de Mitigación de Riesgos: La alta capacidad discriminativa del modelo permite a la gerencia diseñar estrategias efectivas de mitigación de riesgos, enfocándose en áreas identificadas como más vulnerables a fraudes.
- Mejora Continua: Con un ciclo de monitoreo continuo y actualización del modelo, la gerencia puede asegurarse de que el sistema de detección de fraude evolucione con las nuevas amenazas, manteniendo a la organización un paso adelante de los defraudadores.
- Toma de Decisiones Basadas en Datos: La implementación exitosa del modelo promueve una cultura de toma de decisiones basada en datos, donde las políticas y procedimientos se ajustan dinámicamente en función de los hallazgos del análisis de datos.

CONCLUSIONES Y RECOMENDACIONES

En conclusión, se determinó lo siguiente:

- Random Forest demostró ser el más eficaz de acuerdo con las métricas de evaluación propuestas, es así como se cuenta con una alta precisión (0.98), recall (0.77), y un excelente ROC-AUC score (0.9949), lo que indica que se atribuye una alta capacidad para distinguir entre transacciones fraudulentas y no fraudulentas.

Finalmente, la implementación del modelo Random Forest para la detección de fraudes no solo mejora la capacidad de la organización para identificar actividades fraudulentas, sino que también proporciona una base sólida para decisiones gerenciales informadas y estratégicas. Esta integración facilita una respuesta proactiva y adaptativa a las amenazas de fraude, asegurando la protección de los activos financieros y la confianza de los clientes.

REFERENCIAS

Ali, A., Abd Razak, S., Othman, S. H., Eisa, T. A. E., Al-Dhaqm, A., Nasser, M., Elhassan, T., Elshafie, H., & Saif, A. (2022). Financial Fraud Detection Based on Machine Learning: A Systematic Literature Review. *Applied Sciences 2022, Vol. 12, Page 9637*, *12*(19), 9637. https://doi.org/10.3390/APP12199637

Bao, W., Lianju, N., & Yue, K. (2019). Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. *Expert Systems with Applications*, *128*, 301–315. https://doi.org/10.1016/j.eswa.2019.02.033

Borrero-Tigreros, D., & Bedoya-Leiva, O. F. (2020). Predicción de riesgo crediticio en Colombia usando técnicas de inteligencia artificial. *Revista UIS Ingenierías*, *19*(4), 37–52. https://doi.org/10.18273/REVUIN.V19N4-2020004

Dávila-Morán, R. C., Castillo-Sáenz, R. A., Vargas-Murillo, A. R., Dávila, L. V., García-Huamantumba, E., García-Huamantumba, C. F., Cajas, R. F. P., & Paredes, C. E. G. (2023). Aplicación de Modelos de Aprendizaje Automático en la Detección de Fraudes en Transacciones Financieras. *1*, *2*, 109. https://doi.org/10.56294/DM2023109

Doko, F., Kalajdziski, S., & Mishkovski, I. (2021). Credit Risk Model Based on Central Bank Credit Registry Data. *Journal of Risk and Financial Management 2021, Vol. 14, Page 138*, *14*(3), 138. https://doi.org/10.3390/JRFM14030138

Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, *297*(3), 1178–1192. https://doi.org/10.1016/J.EJOR.2021.06.053

Fraudulent Transactions Data. (n.d.). Retrieved July 11, 2024, from https://www.kaggle.com/datasets/chitwanmanchanda/fraudulent-transactions-data/data

Freire López, J. (2021). *Modelo de clasificación de riesgo crediticio utilizando Random Forest en financiera del Ecuador*. http://localhost:8080/xmlui/handle/123456789/4256

Huang, Y., Zhang, L., Li, Z., Qiu, H., Sun, T., & Wang, X. (2020). *Fintech Credit Risk Assessment for SMEs: Evidence from China*. https://papers.ssrn.com/abstract=3721218

Idbenjra, K., Coussement, K., & De Caigny, A. (2024). Investigating the beneficial impact of segmentation-based modelling for credit scoring. *Decision Support Systems*, *179*. https://doi.org/10.1016/j.dss.2024.114170

Li, Z., Liu, G., & Jiang, C. (2020). Deep Representation Learning with Full Center Loss for Credit Card Fraud Detection. *IEEE Transactions on Computational Social Systems*, *7*(2), 569–579. https://doi.org/10.1109/TCSS.2020.2970805

Madaan, M., Kumar, A., Keshri, C., Jain, R., & Nagrath, P. (2021). Loan default prediction using decision trees and random forest: A comparative study. *IOP Conference Series: Materials Science and Engineering*, *1022*(1), 012042. https://doi.org/10.1088/1757-899X/1022/1/012042

Moreno, Á. A., Casasnovas, R. M., & Monzo Sánchez, C. (2023). *Crimen Financiero. Detección de fraude en tarjetas de crédito aplicando aprendizaje automático*. https://openaccess.uoc.edu/handle/10609/148477

Murugan, M. S., & T, S. K. (2023). Large-scale data-driven financial risk management & analysis using machine learning strategies. *Measurement: Sensors*, *27*, 100756. https://doi.org/10.1016/J.MEASEN.2023.100756

Nami, S., & Shajari, M. (2018). Cost-sensitive payment card fraud detection based on dynamic random forest and k-nearest neighbors. *Expert Systems with Applications*, *110*, 381–392. https://doi.org/10.1016/J.ESWA.2018.06.011

Pang, P. S., Hou, X., & Xia, L. (2021). Borrowers' credit quality scoring model and applications, with default discriminant analysis based on the extreme learning machine. *Technological Forecasting and Social Change*, *165*, 120462. https://doi.org/10.1016/J.TECHFORE.2020.120462

Patil, S., Nemade, V., & Soni, P. K. (2018). Predictive Modelling for Credit Card Fraud Detection Using Data Analytics. *Procedia Computer Science*, *132*, 385–395. https://doi.org/10.1016/J.PROCS.2018.05.199

Sahoo, K., Samal, A. K., Pramanik, J., & Pani, S. K. (2019). Exploratory data analysis using python. *International Journal of Innovative Technology and Exploring Engineering*, *8*(12), 4727–4735. https://doi.org/10.35940/IJITEE.L3591.1081219

Segoviano, L. J. G. (2023). Regresion logistica v/s Arboles de desición en el riesgo crediticio.: Logistic regression v/s Decision trees in credit risk. *RICT Revista de Investigación Científica, Tecnológica e Innovación*, *1*(2), 32–37. https://doi.org/10.2992/RICT.V1I2.21

Trivedi, S. K. (2020). A study on credit scoring modeling with different feature selection and machine learning approaches. *Technology in Society*, *63*, 101413. https://doi.org/10.1016/J.TECHSOC.2020.101413

Verster, T., & Fourie, E. (2023). The Changing Landscape of Financial Credit Risk Models. *International Journal of Financial Studies 2023, Vol. 11, Page 98*, *11*(3), 98. https://doi.org/10.3390/IJFS11030098

Xianyu, Q., & Hai, M. (2023). Research on Default Prediction Model of Corporate Credit Risk Based on Big Data Analysis Algorithm. *Procedia Computer Science*, *221*, 300–307. https://doi.org/10.1016/J.PROCS.2023.07.041

Yu, L., Zhou, R., Chen, R., & Lai, K. K. (2022). Missing Data Preprocessing in Credit Classification: One-Hot Encoding or Imputation? *Emerging Markets Finance and Trade*, *58*(2), 472–482. https://doi.org/10.1080/1540496X.2020.1825935

Zhou, Y., Shen, L., & Ballester, L. (2023). A two-stage credit scoring model based on random forest: Evidence from Chinese small firms. *International Review of Financial Analysis*, *89*, 102755. https://doi.org/10.1016/J.IRFA.2023.102755

ANEXOS

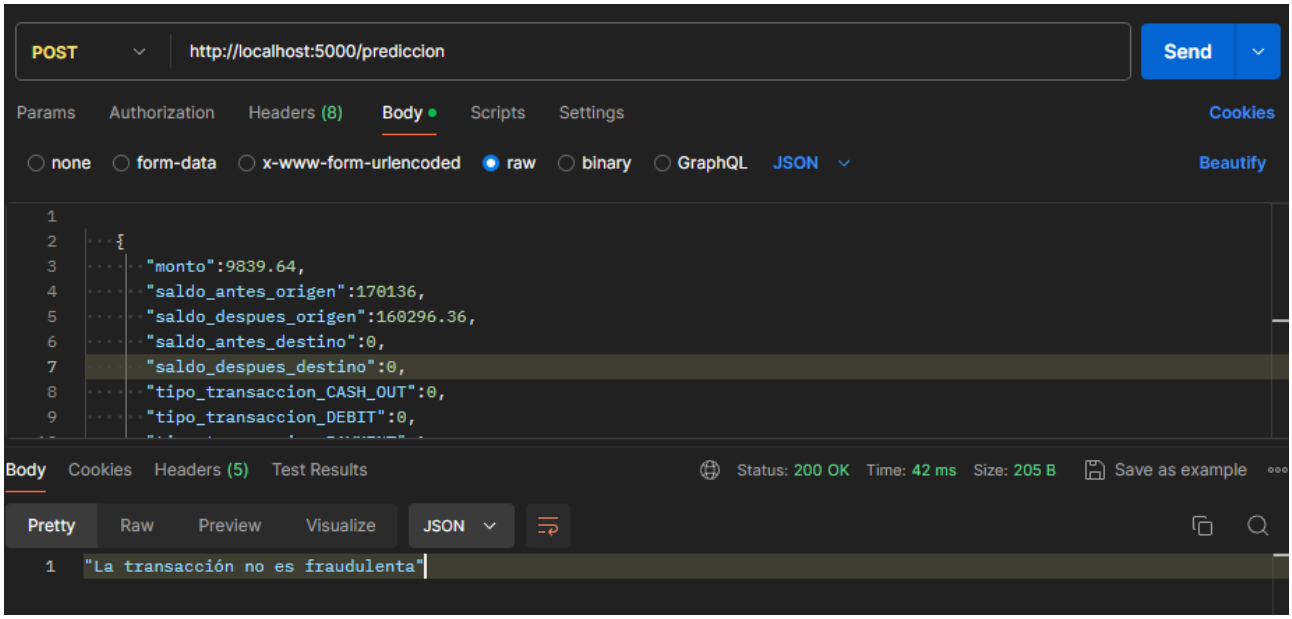
Anexo 1

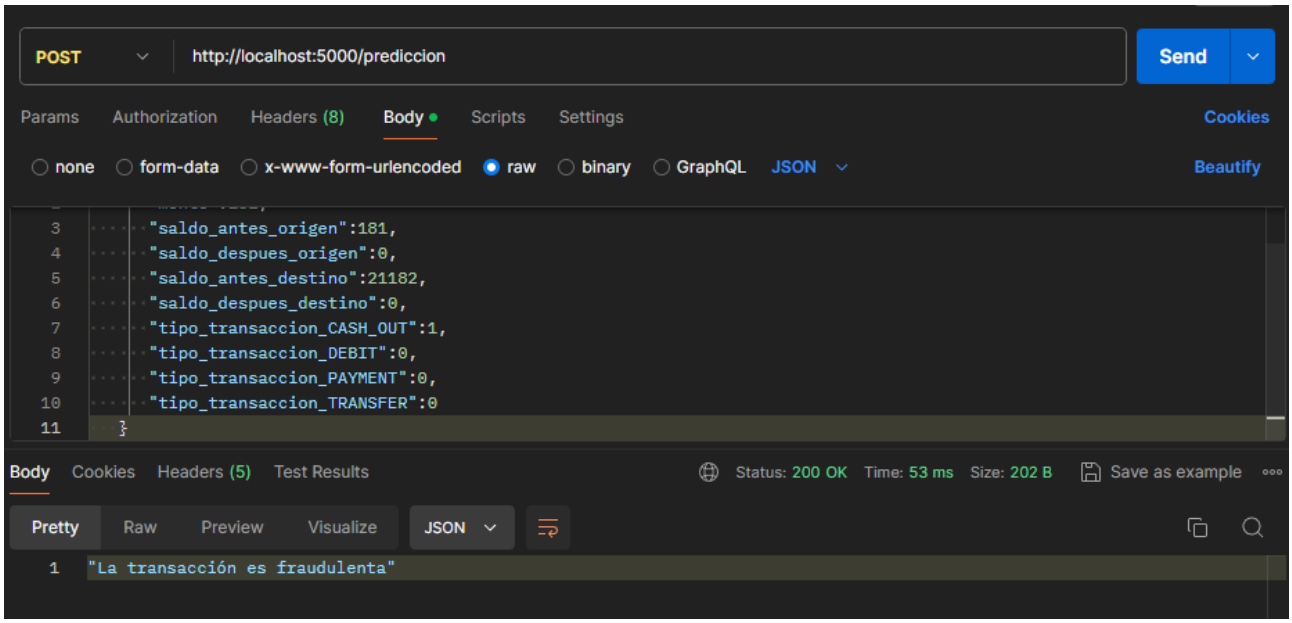
Creación del API, invocando al modelo entrenado

```python
from flask import Flask, request, jsonify
import joblib
import pandas as pd

app = Flask(__name__)

# Cargar el modelo entrenado
modelo = joblib.load('modelo_fraude_financiero.joblib')

@app.route('/prediccion', methods=['POST'])
def prediccion():
    datos = request.get_json(force=True)
    df = pd.DataFrame(datos)

    #Ejecutar el modelo
    predicciones = modelo.predict(df)

    # Obtener el valor de la predicción (suponiendo que solo hay un registro)
    prediccion = predicciones[0]

    # Crear un mensaje basado en la predicción
    if prediccion >= 1:
        resultado = 'La transacción es fraudulenta'
    else:
        resultado = 'La transacción no es fraudulenta'

    return jsonify(resultado)

if __name__ == '__main__':
    app.run(port=5000, debug=False)
```

Anexo 2

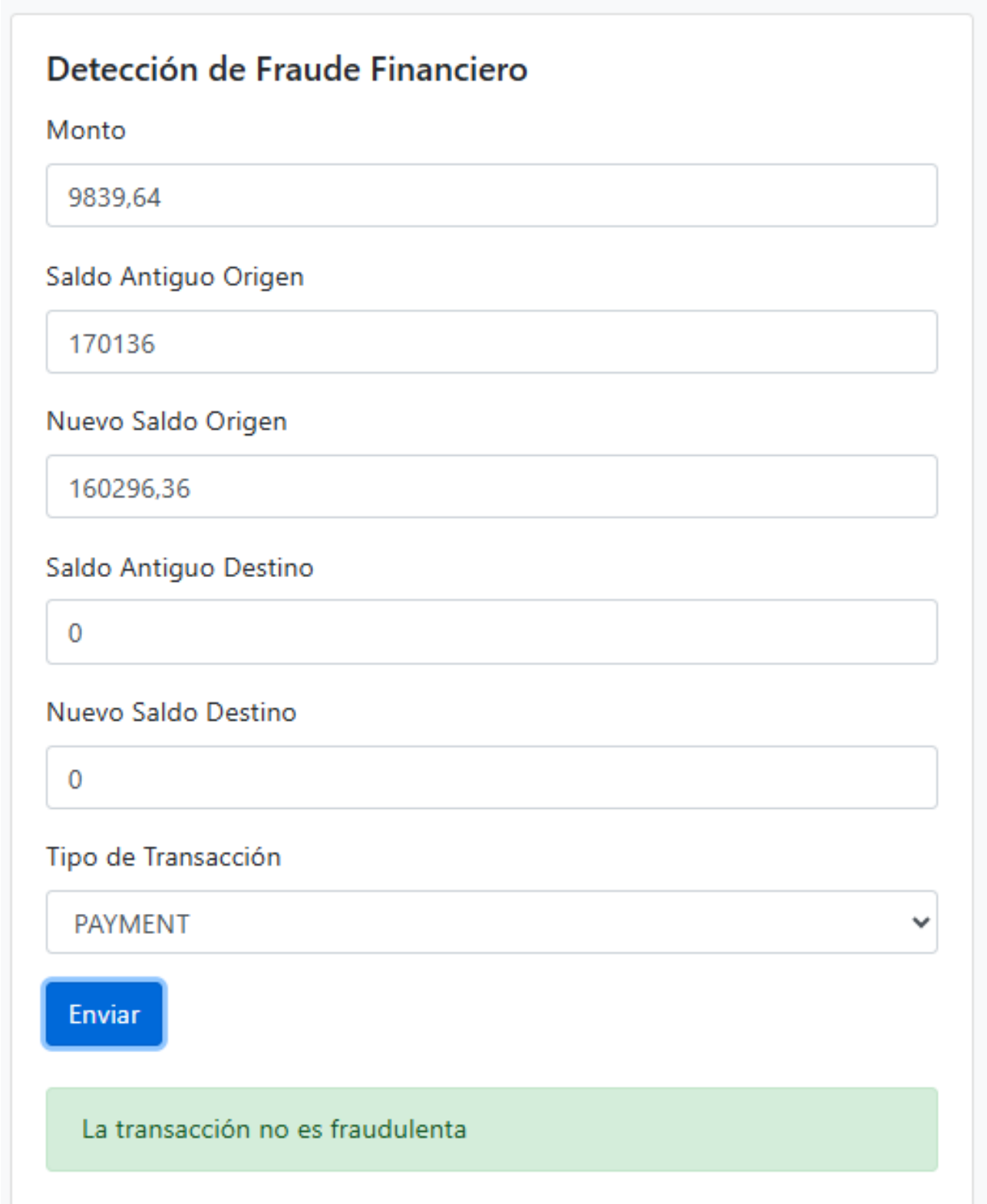
Consumo del API, desde Postman

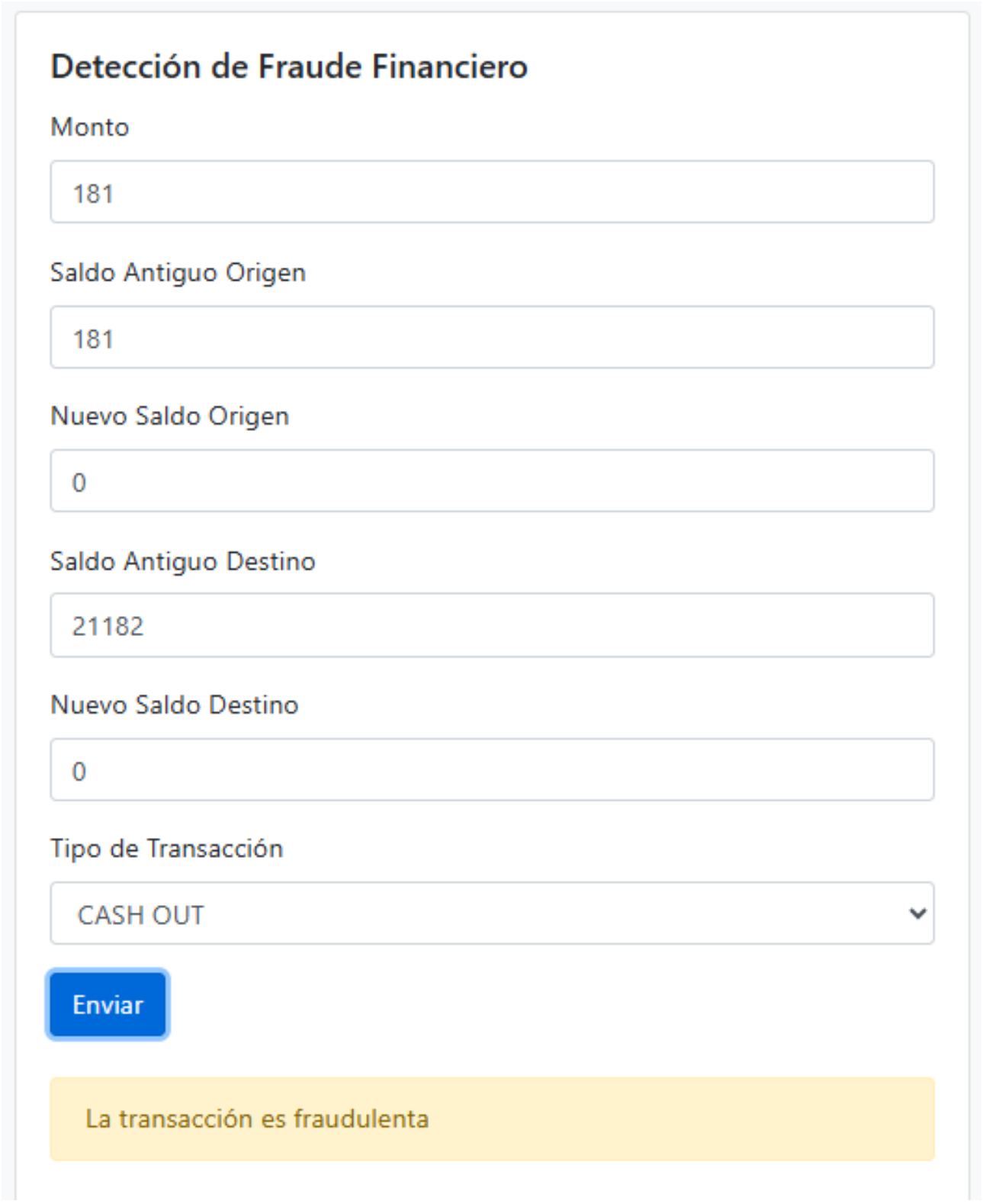




Anexo 3

Integración del API, implementada en la aplicación financiera institucional




Detección de Fraude Financiero
Monto
181
Saldo Antiguo Origen
181
Nuevo Saldo Origen
0
Saldo Antiguo Destino
21182
Nuevo Saldo Destino
0
Tipo de Transacción
CASH OUT
Enviar
La transacción es fraudulenta

Anexo 4

Enlace al Repositorio GitHub con el código del Proyecto

- https://github.com/omadito/Proyecto_Fraude_Financiero