Modelo de Detección de Fraudes en Transacciones Bancarias

González G. Jerónimo, Vélez D. Daniel, Villada C. Juan José

Resumen - El fraude financiero representa una amenaza crítica para instituciones bancarias y plataformas de dinero móvil. Con el crecimiento acelerado de los servicios financieros digitales, surge la necesidad de métodos más robustos y adaptativos para la detección de fraude. Este artículo revisa el estado del arte en técnicas tradicionales y modernas, incluyendo aprendizaje automático, aprendizaje profundo y Graph Neural Networks (GNNs), con un enfoque especial en el uso del dataset sintético PaySim. Se identifican convergencias, contradicciones y vacíos de investigación, destacando la necesidad de modelos explicables, robustos al desbalance de clases y capaces de adaptarse a la evolución del fraude. [1] [2] [3] [5] [8]

Palabras clave — Detección de fraude, Aprendizaje automático, Aprendizaje profundo, Graph Neural Networks, PaySim, Mobile Money. [1] [2] [3] [5]

I. Literature Review & State of the Art

El fraude financiero es un fenómeno crítico en la banca digital y las plataformas de dinero móvil, donde los métodos tradicionales basados en reglas se han quedado cortos frente a patrones cada vez más sofisticados. Históricamente, los sistemas de detección de fraude se fundamentaban en reglas de negocio y estadística clásica. Aunque simples, estos enfoques presentan problemas como alta tasa de falsos positivos, rigidez frente a nuevos esquemas de fraude y baja capacidad para modelar relaciones entre actores [1]. Con el auge del machine learning, se popularizó el uso de algoritmos supervisados como Random Forest, XGBoost y redes neuronales, que mejoran la precisión respecto a los sistemas de reglas [3], [4]. También se han explorado enfoques no supervisados como autoencoders para detección de anomalías. No obstante, la interpretabilidad y el desbalance extremo de clases siguen siendo retos importantes [7].

Más recientemente, las Graph Neural Networks (GNNs) han permitido representar transacciones como grafos de usuarios y comerciantes, capturando patrones colaborativos de fraude. Modelos heterogéneos con mecanismos de atención han mostrado resultados prometedores, destacando su capacidad para aprender relaciones estructurales que los modelos tabulares no logran capturar [2], [5], [6].

La investigación en detección de fraude enfrenta además el reto de acceso a datos reales. PaySim (López-Rojas et al., 2016) es uno de los simuladores más usados, ya que genera transacciones basadas en patrones reales con inyección de fraude. Sin embargo, al ser sintético, no refleja toda la complejidad del fraude en escenarios reales [8].

En síntesis, la literatura coincide en que el machine learning supera a los sistemas de reglas, pero persiste el dilema entre precisión e interpretabilidad. Además, la mayoría de estudios se concentran en tarjetas de crédito, con poca transferencia a dinero móvil. Los datasets sintéticos como PaySim, aunque útiles, no capturan del todo la naturaleza cambiante del fraude. Este trabajo busca contribuir mediante la aplicación de modelos de aprendizaje automático al dataset PaySim, evaluando su desempeño, analizando su interpretabilidad básica y considerando el impacto de los falsos positivos, que constituyen un desafío práctico en la industria para estudios de desarrollo, investigadores y profesionales que buscan innovar en la creación de experiencias interactivas.

II. Research Question & Objectives

La pregunta de investigación que guía este trabajo es: ¿Cómo pueden los modelos de aprendizaje automático detectar eficazmente transacciones fraudulentas en dinero móvil, considerando el fuerte desbalance de clases y la necesidad de interpretabilidad básica?

La motivación radica en que los sistemas actuales presentan problemas de escalabilidad y confiabilidad. Detectar fraudes de manera precisa y explicable puede reducir pérdidas económicas y mejorar la confianza de los usuarios en plataformas de dinero móvil.

Los objetivos del proyecto, definidos bajo el marco SMART, son los siguientes:

- Specific: Entrenar y evaluar un modelo supervisado (Random Forest o XGBoost) sobre PaySim.
- Measurable: Medir desempeño con PR-AUC y F1-score, comparando antes y después del balanceo de clases.
- Achievable: Aplicar una técnica de resampling básica (undersampling o SMOTE) y un análisis de importancia de variables.
- Relevant: Evaluar la capacidad del modelo para identificar fraudes y comprender qué variables más influyen.
- Time-bound: Desarrollar el proyecto en el marco del semestre académico 2025-II.

III. Data & Preliminary Analysis

El conjunto de datos empleado corresponde a PaySim [López-Rojas et al., 2016], un simulador de transacciones financieras móviles basado en datos reales de una compañía africana. Los

datos originales fueron anonimizados y transformados para generar aproximadamente 6.3 millones de registros que representan diferentes tipos de operaciones realizadas por clientes, tanto legítimas como fraudulentas. El dataset contiene 10 variables principales, donde la variable objetivo (y) es isFraud, que indica si una transacción es fraudulenta (1) o legítima (0). Las variables predictoras (X) incluyen: transaccionales como step (tiempo en intervalos de una hora), type (CASH_IN, CASH_OUT, TRANSFER, PAYMENT, DEBIT) y amount (valor de la transacción); de cuentas como oldbalanceOrg, newbalanceOrig, oldbalanceDest y newbalanceDest; y categóricas como nameOrig y nameDest, que representan los identificadores de cliente origen y destino, útiles en modelos basados en grafos, pero no directamente en modelos tabulares. El dataset presenta un alto desbalance de clases, ya que únicamente alrededor del 0.1 % de las transacciones corresponden a fraudes, lo cual refleja la realidad del dominio financiero y constituye un reto central del modelado.

El análisis exploratorio de datos (EDA) muestra que la variable amount exhibe una distribución fuertemente sesgada a la derecha, con la mayoría de transacciones por montos pequeños (por debajo de 1 000 unidades) y una cola larga de transacciones de alto valor. Las transacciones de tipo CASH_OUT y TRANSFER concentran la mayoría de los casos de fraude, mientras que PAYMENT y CASH_IN se asocian casi exclusivamente a operaciones legítimas. Se calculó la matriz de correlación de Pearson para las variables numéricas, observándose correlaciones muy altas entre oldbalanceOrg y newbalanceOrig ($r \approx 0.98$) y entre oldbalanceDest y newbalanceDest ($r \approx 0.97$). Estas relaciones indican que los balances antes y después de cada transacción están fuertemente dependientes, por lo que se generaron variables derivadas como diffOrg = oldbalanceOrg − newbalanceOrig, diffDest = newbalanceDest − oldbalanceDest y ratio = amount / (oldbalanceOrg + 1), con el fin de capturar mejor los movimientos reales de fondos y facilitar la detección de anomalías.

En cuanto a la calidad de los datos, no se identificaron valores faltantes en las variables numéricas ni categóricas. Respecto a valores atípicos, se detectaron montos extremos por encima del percentil 99 que representan transacciones legítimas de gran cuantía, por lo que se optó por conservarlos dado que reflejan operaciones válidas y aportan diversidad al modelo. Durante el análisis se identificó además que la variable isFlaggedFraud se encuentra directamente correlacionada con el objetivo isFraud, por lo que fue excluida del entrenamiento para evitar fugas de información. Asimismo, se decidió emplear únicamente variables disponibles antes de la ejecución de la transacción, como oldbalanceOrg, amount y type, con el fin de simular condiciones realistas de predicción.

El EDA confirma que el problema corresponde a un caso severo de desbalance de clases, lo que hace inadecuado el uso de métricas tradicionales como ROC-AUC. En consecuencia, se utilizarán métricas más informativas, siendo PR-AUC (Precision-Recall Area Under Curve) la métrica principal y F1-score el indicador de equilibrio entre precisión y sensibilidad. Métricas adicionales como recall (maximizar la detección de fraudes) y precision (evitar falsos positivos excesivos) complementarán la evaluación. Los resultados del análisis orientan las siguientes etapas del proyecto: el uso de técnicas de remuestreo (undersampling o SMOTE) para abordar el desbalance, la creación de variables derivadas para mejorar la representatividad de las transacciones y la aplicación de modelos supervisados interpretables que permitan comprender qué factores influyen más en la predicción de fraude.

IV. Materials & Methods

El proceso experimental se desarrolló sobre el conjunto de datos PaySim, un simulador de transacciones financieras móviles basado en datos reales de una empresa africana. El conjunto contiene aproximadamente 6.3 millones de registros y reproduce dinámicas de transferencias legítimas y fraudulentas. La variable objetivo (isFraud) indica si una transacción fue clasificada como fraude (1) o legítima (0). Las variables predictoras incluyen atributos transaccionales (step, type, amount), saldos (oldbalanceOrg, newbalanceOrig, oldbalanceDest, newbalanceDest) y variables categóricas (type).

El preprocesamiento consistió en la eliminación de variables con fuga de información, como isFlaggedFraud, la conversión de tipos de datos y la creación de nuevas variables derivadas:

- **diffOrg**: diferencia entre el saldo inicial y final del emisor.
- **diffDest**: diferencia entre el saldo inicial y final del receptor.
- **ratio**: relación entre el monto de la transacción y el saldo inicial del emisor (amount / (oldbalanceOrg + 1)).

Posteriormente, el conjunto de datos se dividió en entrenamiento y prueba (85 % – 15 %) mediante muestreo estratificado para preservar la distribución de clases. La variable categórica type se codificó mediante *one-hot encoding* antes de aplicar técnicas de balanceo.

Dado el fuerte desbalance de clases (solo 0.1 % de transacciones fraudulentas), se aplicó **SMOTE (Synthetic Minority Oversampling Technique)** para generar ejemplos sintéticos de la clase minoritaria, con una relación final de aproximadamente 1:2 entre fraude y no fraude. Se aplicó además una normalización mediante **MinMaxScaler** para escalar todas las variables en el rango [0,1].

El modelo principal seleccionado fue **XGBoost**, un clasificador de gradiente basado en árboles de decisión. Los hiperparámetros se ajustaron mediante validación cruzada estratificada de 5 pliegues, optimizando el F1-score y el área bajo la curva de precisión-recall (PR-AUC). Como modelo base

se empleó **Random Forest**, entrenado sobre una muestra reducida (15 % del conjunto balanceado) para comparar desempeño y tiempo de entrenamiento.

Las métricas utilizadas fueron **Precision**, **Recall**, **F1-score** y **PR-AUC**, las más apropiadas en contextos de desbalance. Para interpretabilidad, se realizó un análisis de **importancia de variables** usando los pesos reportados por XGBoost.

V. Results

Los resultados cuantitativos obtenidos muestran un desempeño sobresaliente del modelo principal. Tras el preprocesamiento y balanceo, se entrenó XGBoost con un conjunto de aproximadamente 5.4 millones de registros, balanceado a 8.1 millones después del remuestreo. El modelo baseline (Random Forest) se entrenó con el 15 % del conjunto balanceado.

Modelo	F1-score	Precision	Recall	PR-AUC
XGBoost	0.941	0.8911	0.9968	0.9978
RandomForest	0.9812	0.9654	0.9976	—

El modelo XGBoost obtuvo un **F1-score de 0.94**, **precision de 0.89**, **recall de 0.997** y **PR-AUC de 0.9978**, lo que indica un rendimiento casi perfecto en la detección de fraudes. El baseline Random Forest, aunque entrenado sobre un subconjunto, mostró resultados igualmente altos (F1 = 0.98).

Con un soporte de 1 232 casos de fraude en prueba, el modelo XGBoost detectó correctamente aproximadamente 1 228 (FN ≈ 4) y cometió cerca de 137 falsos positivos sobre más de 950 000 transacciones legítimas. Esto representa una tasa de falsos negativos extremadamente baja y un número manejable de falsos positivos, adecuado para aplicaciones donde la prioridad es maximizar la detección de fraude.

El análisis de importancia de XGBoost muestra que las características más influyentes son diffOrg, newbalanceOrig y ratio, seguidas por los tipos de transacción TRANSFER y CASH_OUT. Estas variables capturan cambios anómalos en balances y patrones típicos de fraude, en coherencia con los resultados reportados en la literatura.

VI. Discussion

Los resultados confirman que un modelo supervisado con balanceo y features derivadas puede alcanzar un rendimiento excepcional en detección de fraude financiero. Los valores de recall superiores a 0.99 muestran que el modelo identifica prácticamente todos los fraudes, mientras que la precisión cercana a 0.9 mantiene un bajo nivel de falsas alarmas. Esto cumple plenamente los objetivos planteados en el proyecto y se alinea con la literatura previa sobre PaySim (López-Rojas et al., 2016), que reporta comportamientos similares.

La comparación con el modelo baseline indica que ambos enfoques son altamente efectivos, aunque XGBoost ofrece mayor capacidad de generalización y eficiencia computacional. El análisis de importancia de variables aporta interpretabilidad al proceso: la variable diffOrg —diferencia entre el saldo inicial y final del emisor— domina las decisiones del modelo, seguida por newbalanceOrig y ratio. Este patrón respalda la hipótesis de que los fraudes pueden detectarse observando inconsistencias en balances tras transferencias y retiros.

Análisis de error: los pocos falsos negativos identificados se asocian a transacciones legítimas con características marginales o a fraudes que no alteran significativamente los balances. Los falsos positivos corresponden principalmente a transacciones atípicas pero válidas (retiros o transferencias grandes). Este comportamiento es esperado en sistemas de alto recall y su impacto puede mitigarse mediante ajuste de umbrales y revisión manual en producción.

VII. Conclusions

Este trabajo evaluó modelos supervisados para la detección de fraude en transacciones financieras utilizando el dataset PaySim. Los resultados demuestran que la combinación de preprocesamiento cuidadoso, ingeniería de características (diffOrg, diffDest, ratio) y un clasificador XGBoost balanceado con SMOTE permite detectar fraudes con alta efectividad (F1 ≈ 0.94, PR-AUC ≈ 0.998).

El modelo logró detectar casi la totalidad de los fraudes con una precisión cercana al 90 %, cumpliendo los objetivos definidos de rendimiento y demostrando que enfoques tabulares supervisados son viables para este tipo de tareas.

Cumplimiento de objetivos SMART:

- *Specific/Measurable:* implementación y comparación de modelos con métricas adecuadas al desbalance.
- *Achievable:* aplicación de técnicas de remuestreo (SMOTE) para mejorar el rendimiento.
- *Relevant:* análisis interpretativo de variables más influyentes.
- *Time-bound:* ejecución completa dentro del semestre académico.

Limitaciones: uso de datos simulados y falta de evaluación de variabilidad estadística.
Futuro trabajo: validación en datos reales, exploración de enfoques basados en grafos y análisis temporal, calibración de umbrales para reducir falsos positivos y automatización del retraining con monitoreo de *drift*.

VI. Referencias

[1] S. Selvam and M. Sughasiny, “Smart and Explainable Credit Card Fraud Detection Using XGBoost and SHAP,” J. IoT in Social, Mobile, Analytics, and Cloud, vol. 7, no. 2, pp. 155–169, Jul. 2025.

[2] Q. Sha et al., “Detecting Credit Card Fraud via Heterogeneous Graph Neural Networks with Graph Attention,” arXiv, Jun. 2025.

[3] “Enhancing Fraud Detection in Credit Card Transactions: A Comparative Study of Machine Learning Models,” Computational Economics, 2025.

[4] “Application of Machine Learning Models for Fraud Detection in Synthetic Mobile Financial Transactions,” JITK, 2024.

[5] “Financial fraud detection using graph neural networks: A systematic review,” Expert Systems with Applications, vol. 240, Apr. 2024.

[6] “Graph neural network for fraud detection via context encoding and adaptive aggregation,” Expert Systems with Applications, vol. 261, Feb. 2025.

[7] “Interpretable Ensemble Learning Models for Credit Card Fraud Detection,” Preprints.org, 2025.

[8] “Predicting mobile money transaction fraud using machine learning algorithms,” Applied AI Letters, 2023.