# Identificación de Fraudes en Transacciones Electrónicas mediante Aprendizaje Automático Explicable

Silvia Vázquez
Universidad Americana
silvia.vazquez@americana.edu.py
Asunción-Paraguay

***Abstract*—Fraud identification in electronic payment systems remains a significant challenge for progress in anti-fraud efforts in electronic payment systems. This paper proposes a predictive model based on XGBoost in combination with Explainable Artificial Intelligence (XAI) techniques such as SHAP, which offers clarity and high accuracy in detecting fraudulent transactions. A public dataset was used, and all steps were carried out with open-source tools (Python, scikit-learn, XGBoost, SHAP, Google Colab), ensuring reproducibility and auditability of the experiments.**

***Keywords*—Fraud Detection; Explainable AI; SHAP; XGBoost; Open Source**

***Resumen*— La identificación de fraudes en sistemas de pago electrónico, sigue representando un reto significativo para el progreso antifraude de los sistemas de pago electrónico. En este trabajo se propone un modelo predictivo basado en XGBoost en combinación con técnicas de Inteligencia Artificial Explicable (XAI), como SHAP, que ofrece claridad y gran exactitud en la detección de transacciones fraudulentas. Se empleó un conjunto de datos públicos, y todas las etapas se llevaron a cabo con herramientas de código abierto (Python, scikit-learn, XGBoost, SHAP, Google Colab) garantizando así su reproducibilidad y auditoría de los experimentos.**

***Palabras claves* —Identificación de fraudes; Inteligencia Artificial Explicable; SHAP; XGBoost; Software Libre**

## I. INTRODUCCIÓN

Detectar operaciones fraudulentas en sistemas de pago electrónicos es uno de los principales desafíos del sector financiero. Las consecuencias económicas, legales y reputacionales derivadas de estas actividades han impulsado el desarrollo de modelos predictivos cada vez más precisos [1]. En los últimos años, las técnicas de aprendizaje automático han demostrado eficacia comprobada para identificar patrones irregulares en grandes volúmenes de datos transaccionales, contribuyendo de manera significativa a la prevención del fraude financiero [2].

Sin embargo, la creciente complejidad de estos modelos ha dificultado su comprensión por parte de usuarios finales y auditores. En este contexto, la Inteligencia Artificial Explicable (XAI) surge como una metodología orientada a proporcionar transparencia a los sistemas predictivos, facilitando la interpretación de las razones que sustentan cada decisión del modelo [3].

Este estudio presenta resultados preliminares sobre la implementación y validación de un modelo predictivo con capacidad explicativa, desarrollado con el objetivo de evaluar su desempeño y potencial de interpretabilidad en la detección de fraudes electrónicos.

## II. DESARROLLO

### *A. Marco Teórico*

El campo del aprendizaje automático para detectar fraude financiero se encuentra actualmente en rápida expansión, motivado por la demanda de gestionar grandes cantidades de transacciones y la creciente complejidad de los atacantes. Los modelos supervisados, como la regresión logística, los árboles de decisión, Random Forest y XGBoost [4], se han utilizado extensamente para categorizar transacciones en legítimas o fraudulentas. Sin embargo, la mayoría de estos modelos operan como cajas negras, es decir, proporcionan resultados sin detallar de manera precisa cómo se adoptó la decisión, lo que podría obstaculizar la confianza y la adopción en contextos regulados o críticos.

Para vencer estas restricciones, la Inteligencia Artificial Explicable (XAI) se ha desarrollado como una disciplina que aspira a proporcionar a los modelos predictivos mecanismos que faciliten la explicación de sus decisiones.

### *B. Metodología*

El estudio se desarrolló sobre el conjunto de datos público Credit Card Fraud Detection Dataset de Kaggle, que contiene 284 807 transacciones con tarjetas de crédito, de las cuales 492 corresponden a fraudes. Este dataset se caracteriza por un marcado desbalance de clases y es ampliamente utilizado como referencia para investigaciones en detección de fraude financiero [2].

El trabajo siguió un pipeline reproducible compuesto por cuatro etapas principales:
(a) preprocesamiento de datos,
(b) segmentación del conjunto de datos,
(c) entrenamiento del modelo de aprendizaje automático, y
(d) evaluación e interpretación de resultados mediante técnicas de explicabilidad.

#### *1) Preprocesamiento de datos*

Se verificaron tipos de datos, rangos y consistencia, descartando registros duplicados o inconsistentes. No se identificaron valores nulos en la versión utilizada. La variable objetivo fue Class, con codificación binaria (0 = transacción legítima, 1 = fraude). Las variables predictoras correspondieron a los componentes anónimos V1–V28, junto con Amount y Time.
Para evitar fuga temporal, la variable Time se excluyó del modelado, dado que su inclusión podría introducir dependencias no generalizables.

El algoritmo XGBoost no requiere normalización ni escalado de características; por ello, las variables se mantuvieron en su rango original. Asimismo, se evaluó la colinealidad y la varianza de los predictores, sin observarse redundancias significativas que justificaran su eliminación.

#### *2) Segmentación del conjunto de datos*

Se realizó una partición del conjunto total en 70 % para entrenamiento y 30 % para prueba, conservando la proporción original de clases mediante estratificación. Este procedimiento siguió un esquema de evaluación hold-out, con el fin de garantizar independencia entre las fases de entrenamiento y validación.

#### *3) Entrenamiento del modelo*

El modelo se entrenó utilizando el algoritmo Extreme Gradient Boosting (XGBoost) debido a su alto rendimiento en problemas con datos desbalanceados [4]. Los hiperparámetros predeterminados del modelo fueron mantenidos sin ajuste fino, priorizando la reproducibilidad.
El entrenamiento se ejecutó en el entorno Google Colab, empleando bibliotecas de código abierto como Python, scikit-learn, XGBoost y SHAP, garantizando la trazabilidad y apertura del experimento.

#### *4) Evaluación e interpretabilidad*

El desempeño se evaluó sobre el conjunto de prueba utilizando métricas de precisión, recall y F1-score, además de la matriz de confusión (Tabla I).
Para la interpretabilidad, se aplicó la técnica SHAP (SHapley Additive exPlanations), que descompone la predicción de cada observación en las contribuciones individuales de las variables, basándose en la teoría de juegos de Shapley [3].

Se generaron dos tipos de visualizaciones explicativas:

- **Figura 1.** *Importancia global de las variables según SHAP*, que resume la influencia de cada predictor en las decisiones del modelo.
- **Figura 2.** *Explicación individual de una transacción mediante SHAP*, que ilustra cómo la combinación de variables impulsa la clasificación hacia “fraude” o “legítima”.

Ambas figuras se discuten posteriormente en la sección de Resultados y Discusión, donde se analizan los hallazgos obtenidos y la relevancia de las variables más influyentes.

## III. RESULTADOS Y DISCUSIÓN

El modelo XGBoost alcanzó resultados satisfactorios en la detección de fraudes utilizando el esquema de evaluación hold-out.
A pesar del marcado desbalance de clases y de no haberse aplicado técnicas de balanceo, el desempeño general fue adecuado, evidenciando la capacidad del modelo para distinguir entre transacciones legítimas y fraudulentas.

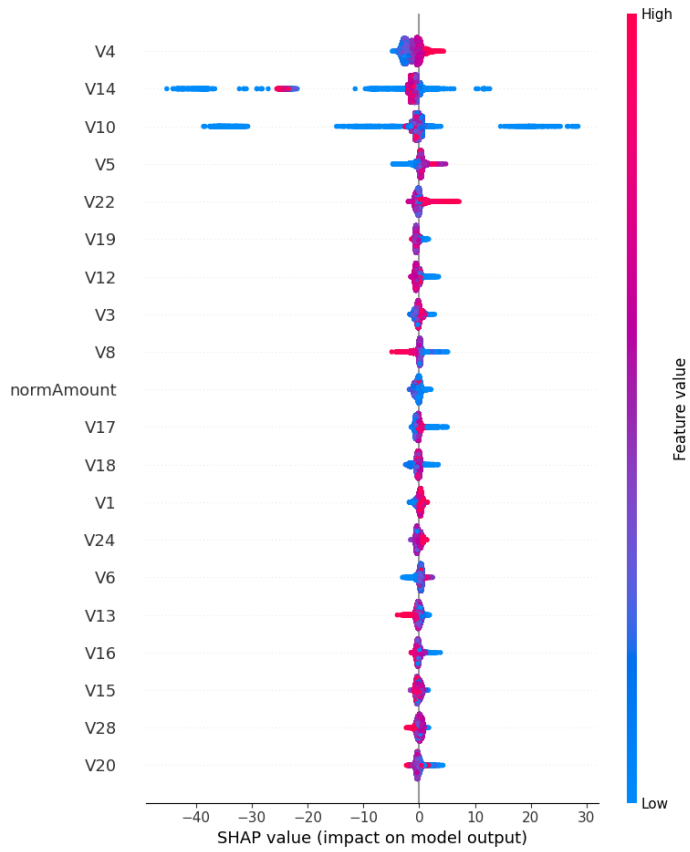
La Tabla I resume las métricas de evaluación obtenidas sobre el conjunto de prueba. Se observa que el modelo presenta una precisión del 88 %, un recall del 76 % y un F1-score del 81 % en la clase minoritaria (fraude), lo que demuestra un equilibrio razonable entre la detección correcta de fraudes y la minimización de falsos positivos.

Tabla I. Métricas de desempeño del modelo XGBoost

| Métrica | Valor |
|---|---|
| Precisión (fraude) | 88% |
| Recall (fraude) | 76% |
| F1-Score (fraude) | 81% |

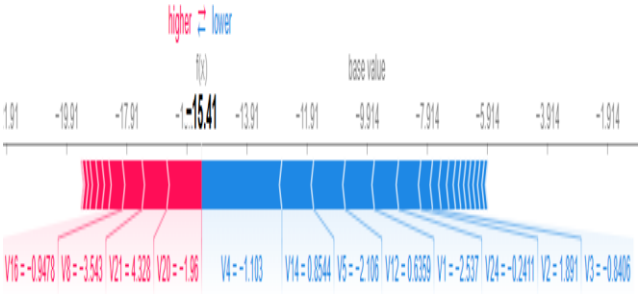
La Figura 1 presenta la importancia global de las variables según SHAP, donde los predictores V14, V17 y Amount se identificaron como los de mayor contribución en las decisiones del modelo. El color de cada punto representa el valor de la variable, permitiendo visualizar cómo los valores altos o bajos influyen en la probabilidad de fraude.

Figura 1. Importancia global de las variables según SHAP.




Por su parte, la Figura 2 muestra una explicación individual de una transacción a través del diagrama tipo force plot, que evidencia cómo las variables con mayor impacto (en rojo) impulsan la predicción hacia "fraude", mientras que las de menor impacto (en azul) la alejan de esa clasificación.
Este tipo de visualización permite auditar y comprender el comportamiento del modelo en casos concretos.

*Figura 2. Explicación individual de una transacción mediante SHAP.*




El análisis visual obtenido mediante SHAP confirma que el modelo no solo logra un rendimiento aceptable, sino que también ofrece transparencia e interpretabilidad, cualidades fundamentales para su adopción en entornos financieros regulados.
La posibilidad de identificar las variables más influyentes y explicar cada decisión refuerza la trazabilidad y confianza en el sistema predictivo.

En conjunto, los resultados muestran que la combinación de XGBoost con técnicas de inteligencia artificial explicable constituye un enfoque viable para el diagnóstico y la auditoría de fraudes electrónicos, aun en presencia de un fuerte desbalance de clases.

## IV. CONCLUSION

El estudio demostró la viabilidad de aplicar modelos de aprendizaje automático, específicamente XGBoost, combinados con técnicas de inteligencia artificial explicable (XAI), para la detección de transacciones fraudulentas en sistemas de pago electrónico. Los resultados obtenidos muestran que el modelo logra un equilibrio adecuado entre precisión (88 %) y capacidad de detección (recall del 76 %), incluso sin la utilización de técnicas avanzadas de balanceo de clases.

La incorporación de SHAP permitió analizar la contribución individual de las variables, otorgando transparencia al proceso de decisión y facilitando la trazabilidad del modelo. Este enfoque posibilita que las instituciones financieras puedan auditar y justificar las predicciones automatizadas, fortaleciendo la confianza en sistemas de apoyo a la toma de decisiones.

En futuras etapas se prevé la integración de estrategias de balanceo de clases y el ajuste de hiperparámetros, además de la evaluación de otros algoritmos y conjuntos de datos.De esta manera, se busca consolidar un sistema predictivo más robustas, interpretable y aplicable en entornos reales de prevención de fraude financiero.


## AGRADECIMIENTOS

Se agradece el apoyo brindado por el cuerpo docente, los equipos de investigación y el personal técnico que colaboraron en las distintas etapas del proyecto. Su orientación metodológica, el acceso a recursos de laboratorio y el intercambio de conocimientos fueron fundamentales para la correcta ejecución de los experimentos y la validación de resultados. Asimismo, se reconoce la contribución institucional de la Universidad Americana, que facilitó las condiciones necesarias para el desarrollo de este estudio.


## REFERENCIAS


[1] A. Dal Pozzolo, O. Caelen, R. Johnson, and G. Bontempi, "Calibrating Probability with Undersampling for Unbalanced Classification," *Proc. IEEE Symposium Series on Computational Intelligence*, pp. 159–166, 2015. DOI: 10.1109/SSCI.2015.33

[2] Kaggle, "Credit Card Fraud Detection Dataset," 2018. [Online]. Available: https://www.kaggle.com/mlg-ulb/creditcardfraud. [Accessed: Jul. 8, 2025].

[3] S. M. Lundberg and S.-I. Lee,"A Unified Approach to

Interpreting Model Predictions,"*Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 4765–4774, 2017.

[4] T. Chen and C. Guestrin,"XGBoost: A Scalable Tree Boosting System,"*Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.DOI: 10.1145/2939672.2939785

[5] M. Carcillo, Y.-A. Le Borgne, O. Caelen, Y. Kessaci, F. Oblé, and G. Bontempi,"Combining Unsupervised and Supervised Learning in Credit Card Fraud Detection," *Information Sciences*, vol. 557, pp. 317–331, Sept. 2021. DOI: 10.1016/j.ins.2020.12.015