Innovación Ciencia y Tecnología /enero –junio 2026/ Vol. 4, - No. 1
Doi:10.70577/innovascit.v4i1.105

Análisis de interpretabilidad mediante técnicas XAI en modelos de Machine Learning para la detección de fraudes financieros

Interpretability analysis using XAI techniques in Machine Learning models for the detection of financial fraud

RESUMEN

La detección de fraudes financieros mediante modelos de Machine Learning es una estrategia eficaz; sin embargo, la complejidad de estos algoritmos ha incrementado la opacidad en las decisiones automatizadas, representando un desafío crítico en sectores regulados que exigen auditabilidad y confianza. Este estudio experimental analiza la interpretabilidade de modelos aplicados a la detección de fraudes en tarjetas de crédito utilizando datos públicos (IEEE-CIS). Siguiendo la metodología CRISP-DM, se implementaron los modelos Random Forest y XGBoost, evaluando el desempeño mediante métricas robustas al desequilibrio extremo, priorizando la Precisão Média (PR-AUC) y el F1-score. Posteriormente, se aplicaron las técnicas SHAP y LIME para generar explicações post-hoc a nivel global y local. Los resultados evidencian que el XGBoost obtuvo el desempeño superior, con un PR-AUC de 0.884 ± 0.012 y Recall de 0.821, mitigando eficazmente el desequilibrio de la clase. Las técnicas XAI permitieron identificar las variables más influyentes globalmente y proporcionar justificações locales consistentes para instâncias individuales. No obstante, la explicación de variables anonimizadas por PCA limitó la interpretabilidade semántica a una validación puramente técnica. En conclusión, el estudio aporta evidencia empírica de que es viável alcanzar un equilibrio entre desempeño preditivo e interpretabilidade técnica, contribuyendo al desarrollo de sistemas de detección de fraudes más transparentes y relevantes para ambientes financieros regulados.

PALABRAS CLAVE: Machine Learning; fraude financiero; inteligencia artificial explicable; interpretabilidad; XAI

ABSTRACT

Telemedicine has become established as the use of information and communication technologies for diagnosis, treatment, follow-up, and, especially, remote mental health education. In young adults (18–29 years), a population with a high prevalence of disorders such as depression and anxiety, this modality overcomes geographic, economic, and stigma barriers, offering opportunities for early prevention and empowerment in psychological self-care. Objective: To determine the reach and effectiveness of telemedicine-based mental health educational interventions for young adults by analyzing their outcomes in knowledge, attitudes, symptoms, and acceptance. Methodology: A systematic review was conducted of quantitative (RCTs, quasi-experimental, before-and-after), qualitative, and mixed studies that evaluated telemedicine-based mental health educational interventions in young adults (18–25 years). Databases (PubMed, Scopus, Web of Science, Google Scholar) were explored without time limit, in English and Spanish. Interventions with an explicit educational component by videoconference, apps, web platforms or messaging were included, and outcomes such as knowledge, attitudes, help-seeking, symptoms and acceptability were considered. Results: The reviewed studies show that telephone and face-to-face transdiagnostic interventions (CETA) reduce internal and external symptoms in young adults, in addition personalized.

KEYWORDS: Telemedicine; Telepsychology; Mental health education; Young adults; Digital interventions; Systematic review.

EDUCATECH

Recepción: 11/04/2026
Aceptación: 22/04/2026
Publicación: 30/06/2026

AUTOR/ES

Macias Veliz Carlos Alfredo
Zambrano Montenegro David Fernando

cmacias5170@utm.edu.ec
deivizamm@gmail.com

Universidad Tècnica de Manabí
Universidad Tècnica de Manabí

Portoviejo – Ecuador
Portoviejo – Ecuador

CITACIÓN:
Macias, C. & Zambrano, D. (2026). Análisis de interpretabilidad mediante técnicas XAI en modelos de Machine Learning para la detección de fraudes financieros. Revista InnovaSciT. *4 (1,). p. 139 - 157.*

INTRODUCCIÓN

La transformación digital del sector financiero ha impulsado una adopción acelerada de sistemas basados en Machine Learning (ML) para la detección de fraudes en transacciones electrónicas, pagos digitales, banca en línea y plataformas fintech. Estos modelos han demostrado una alta capacidad para identificar patrones complejos y anomalías en grandes volúmenes de datos financieros, superando ampliamente a los métodos tradicionales basados en reglas estáticas (Ahmed et al., 2023). Sin embargo, este aumento en el rendimiento predictivo ha venido acompañado de un problema crítico: la falta de interpretabilidad y transparencia de los modelos de aprendizaje automático más avanzados.

Los modelos comúnmente utilizados en la detección de fraudes —como Random Forest, XGBoost, redes neuronales profundas y modelos basados en grafos— suelen operar como sistemas de caja negra, donde el proceso interno que conduce a una decisión resulta opaco para analistas, auditores y reguladores (Gramegna & Giudici, 2021). Esta opacidad representa un riesgo significativo en contextos financieros, donde una predicción errónea puede derivar en el bloqueo injustificado de transacciones legítimas, afectando derechos fundamentales de los usuarios y generando impactos económicos y reputacionales para las instituciones (Misneviciute & Kovacevic, 2022).

En respuesta a estos desafíos, la Inteligencia Artificial Explicable (Explainable Artificial Intelligence, XAI) ha emergido como un campo de investigación clave orientado a proporcionar mecanismos que permitan comprender, interpretar y justificar las decisiones de los modelos de ML. Según Rahman et al. (2024), la XAI no solo mejora la comprensión técnica de los modelos, sino que también fortalece la confianza de los usuarios, facilita la auditoría de sistemas automatizados y apoya el cumplimiento de normativas regulatorias cada vez más exigentes en el sector financiero.

Diversos estudios recientes destacan que la explicabilidad se ha convertido en un requisito esencial para la adopción responsable de la inteligencia artificial en dominios de alto riesgo, como la banca y los servicios financieros digitales. Ahmed et al. (2023) señalan que los modelos explicables permiten identificar variables clave asociadas al fraude, facilitando la validación de decisiones algorítmicas por parte de expertos humanos. De manera similar, Grado-Caballero et al. (2022) argumentan que la ausencia de interpretabilidad limita la capacidad de los analistas para detectar sesgos algorítmicos y errores sistemáticos, lo que compromete la equidad y la transparencia del sistema.

Entre las técnicas XAI más utilizadas en la literatura reciente se encuentran los métodos post-hoc, especialmente SHAP (SHapley Additive exPlanations) y LIME (Local Interpretable Model-agnostic Explanations). SHAP, basado en la teoría de juegos cooperativos, permite cuantificar la contribución de cada variable a una predicción individual o global, proporcionando explicaciones consistentes y matemáticamente fundamentadas (Zheng et al.,

2023). Por su parte, LIME genera modelos locales aproximados que facilitan la interpretación de decisiones específicas en modelos complejos (Gan et al., 2023). Estas técnicas han sido ampliamente aplicadas en la detección de fraudes financieros debido a su flexibilidad y capacidad para trabajar con modelos de alto rendimiento.

No obstante, a pesar de su creciente adopción, la literatura evidencia importantes limitaciones en la evaluación de la explicabilidad. Moscato et al. (2022) advierten que muchas explicaciones generadas pueden ser inestables ante pequeñas variaciones en los datos, lo que reduce su fiabilidad para la toma de decisiones críticas. Asimismo, Sosa y Adhikari (2024) señalan que la mayoría de los estudios priorizan métricas de rendimiento predictivo —como precisión, recall o F1-score— mientras relegan la evaluación cuantitativa de la calidad de las explicaciones, generando una brecha entre desempeño y transparencia.

Desde una perspectiva regulatoria y ética, la interpretabilidad adquiere una relevancia aún mayor. Fernandez y Smith (2021) destacan que los organismos reguladores financieros demandan sistemas auditables y explicables que permitan justificar decisiones automatizadas, especialmente en escenarios donde estas afectan el acceso a servicios financieros. En el contexto latinoamericano, Vasquez-Arnez y Cruz (2023) subrayan que la falta de explicabilidad en modelos de fraude puede profundizar la exclusión financiera y reproducir sesgos estructurales en mercados emergentes, donde los historiales crediticios suelen ser incompletos o desiguales.

En Ecuador, el incremento de transacciones digitales y la adopción de soluciones basadas en ML para la seguridad financiera han generado un escenario similar. Larrea-Vizuete y Castro (2022) evidencian que, si bien las instituciones financieras han incorporado modelos de alta complejidad para combatir el fraude, persisten limitaciones en cuanto a la transparencia de sus decisiones, lo que dificulta la aceptación pública y la supervisión efectiva de estos sistemas. En este contexto, la aplicación de técnicas XAI sobre conjuntos de datos públicos representa una oportunidad para generar conocimiento técnico replicable y adaptado a la realidad nacional.

A partir de este panorama, se vuelve evidente la necesidad de realizar un análisis sistemático y experimental de la interpretabilidad de modelos de Machine Learning aplicados a la detección de fraudes financieros. Evaluar cómo las técnicas XAI permiten revelar patrones de decisión, identificar variables críticas y mejorar la comprensión de los modelos no solo contribuye al avance científico, sino que también aporta soluciones prácticas para el desarrollo de sistemas financieros más transparentes, confiables y éticamente responsables (Djeumou & Chiuso, 2023).

En consecuencia, el presente estudio se orienta a analizar la interpretabilidad de modelos de ML mediante técnicas de Inteligencia Artificial Explicable, utilizando un conjunto de datos público de fraudes financieros. A través de la aplicación de métodos como SHAP y

LIME, se busca aportar evidencia empírica sobre su utilidad para mejorar la transparencia, apoyar la auditoría de modelos y fortalecer la confianza en sistemas automatizados de detección de fraude, contribuyendo así al desarrollo de una inteligencia artificial financiera alineada con principios de responsabilidad, explicabilidad y justicia algorítmica.

MÉTODOS MATERIALES

Este trabajo corresponde a un estudio experimental computacional aplicado, orientado a comparar el desempeño predictivo y la explicabilidad post-hoc de modelos de clasificación para detección de fraude. La evaluación se realizó con un protocolo reproducible basado en partición estratificada y validación cruzada, fijando una semilla aleatoria para controlar la variabilidad estocástica: se utilizó [train/test split = 80%/20%] con stratify y random_state=42. Para el ajuste de hiperparámetros se empleó RandomizedSearchCV con validación cruzada estratificada StratifiedKFold(n_splits=5, shuffle=True, random_state=42), optimizando la métrica Average Precision (PR-AUC).

La selección de Average Precision (PR-AUC) como métrica de optimización y evaluación primaria se fundamenta en la naturaleza intrínsecamente desbalanceada de los conjuntos de datos de fraude financiero. A diferencia del área bajo la curva ROC (ROC-AUC), la cual puede proporcionar una visión optimista y sesgada cuando la clase positiva es minoritaria, la métrica PR-AUC se centra exclusivamente en el desempeño del modelo respecto a la clase crítica (fraude).

Al integrar la precisión y el recall en un espectro de umbrales de decisión, el PR-AUC penaliza de forma más estricta los falsos positivos en entornos donde el costo de inspección manual es elevado, garantizando que las predicciones positivas sean altamente confiables sin sacrificar la capacidad de detección del sistema. Esta aproximación asegura un balance robusto entre la eficacia operativa y el rigor estadístico necesario para modelos de clasificación en dominios de alta criticidad.

El estudio siguió CRISP-DM como marco de trabajo para organizar las fases de comprensión del problema, comprensión de datos, preparación de datos, modelado y evaluación. Se documentaron explícitamente las decisiones que afectan reproducibilidad (partición de datos, semilla, preprocesamiento sin fuga de información, balanceo aplicado solo en entrenamiento, hiperparámetros finales y versiones de librerías), integrando técnicas XAI en la etapa de análisis post-entrenamiento.

Para el desarrollo experimental, se seleccionó el conjunto de datos IEEE-CIS Fraud Detection (2019), desarrollado por la IEEE Computational Intelligence Society en colaboración con Vesta Corporation. Este repositorio, que constituye un estándar contemporáneo (benchmark) en la literatura, comprende más de 590,000 transacciones con una tasa de fraude aproximada del 3.5%, representando un escenario de desbalance de clases característico de entornos financieros reales. El uso de este dataset facilita la reproducibilidad del estudio y

elimina las restricciones éticas asociadas al manejo de información financiera sensible, permitiendo una comparación directa con investigaciones recientes de alto impacto (IEEE-CIS, 2019).

El conjunto de datos integra variables críticas como TransactionDT (distancia temporal), TransactionAmt e identificadores de identidad y producto, proporcionando un marco robusto para el análisis de explicabilidad post-hoc. No obstante, es importante señalar que el dataset contiene dimensiones anonimizadas (específicamente del bloque V1 al V339) cuya interpretabilidad semántica directa es limitada. Bajo estas condiciones, el estudio se desarrolla manteniendo un entorno metodológicamente controlado que equilibra la complejidad de los datos reales con el rigor del análisis computacional.

La evaluación se realizó mediante un protocolo estrictamente reproducible basado en partición estratificada y validación cruzada, fijando una semilla aleatoria para controlar la variabilidad estocástica. Se utilizó un esquema de [train/test split = 80%/20%] con el parámetro stratify activado y un random_state=42. Para el ajuste fino de hiperparámetros, se empleó RandomizedSearchCV con validación cruzada estratificada StratifiedKFold(n_splits=5, shuffle=True, random_state=42). Dada la naturaleza desbalanceada de la muestra, se seleccionó la métrica Average Precision (PR-AUC) como función objetivo de optimización, garantizando una evaluación centrada en la precisión y el recall de la clase minoritaria.

La fase de preprocesamiento fue fundamental para garantizar la calidad de los modelos entrenados y la fiabilidad de las explicaciones generadas. Tras un análisis exploratorio de datos (EDA), se aplicaron las siguientes técnicas siguiendo un protocolo de prevención de fuga de información (data leakage):

- Limpieza y Tratamiento de Valores Faltantes: Se verificó la presencia de valores nulos, identificando que diversas variables de los bloques V (características de ingeniería) e Identidad presentaban una tasa de ausencia superior al 40%. Siguiendo criterios de calidad de datos, se eliminaron las columnas con más del 50% de valores faltantes. Para el resto de los atributos, se aplicó una imputación por la mediana en variables numéricas y se asignó una categoría específica denominada 'Unknown' para las variables categóricas. No se identificaron registros duplicados significativos, asegurando la unicidad de las transacciones analizadas.
- Codificación de Variables Categóricas: Dado que el dataset IEEE-CIS incluye atributos no numéricos críticos (e.g., ProductCD, card4, P_emaildomain), se implementó un esquema de Label Encoding para variables con alta cardinalidad y One-Hot Encoding para aquellas con categorías limitadas. Este proceso permitió transformar la información cualitativa en un formato matricial apto para algoritmos de aprendizaje automático, manteniendo la integridad semántica necesaria para la fase de explicabilidad.

- Escalado de Características: Debido a que variables como TransactionAmt presentan distribuciones con un sesgo positivo extremo y presencia de valores atípicos (outliers), se aplicó una transformación logarítmica seguida de un RobustScaler. Este método de escalado es preferible al estándar en detección de fraude, ya que utiliza el rango intercuartílico, minimizando la influencia de transacciones anómalas de gran magnitud. Es imperativo señalar que el escalador se ajustó exclusivamente con el conjunto de entrenamiento y se aplicó posteriormente al conjunto de prueba para evitar cualquier forma de data leakage.
- Tratamiento del Desbalance de Clases: Ante el marcado desbalance de la muestra (aproximadamente 3.5% de fraude), se implementó la técnica de sobremuestreo sintético SMOTE (Synthetic Minority Over-sampling Technique). Se configuró una estrategia de muestreo sampling_strategy=0.1 para fortalecer la representación de la clase minoritaria de forma controlada. Para garantizar la validez metodológica, esta técnica se aplicó únicamente sobre el conjunto de entrenamiento tras realizar la partición estratificada, asegurando que la evaluación en el conjunto de prueba se realizara sobre una distribución de datos reales no alterados.

En la fase de modelado se implementaron dos algoritmos de aprendizaje supervisado, seleccionados por su alto desempeño demostrado en datos tabulares financieros y su capacidad para capturar relaciones no lineales complejas:

- Random Forest (RF): Un método de ensamblaje (bagging) basado en múltiples árboles de decisión. Para mitigar el desbalance intrínseco del dataset, se configuró el parámetro class_weight='balanced_subsample', el cual ajusta los pesos de las clases basándose en la frecuencia de los datos en cada árbol individual durante el entrenamiento.
- XGBoost (Extreme Gradient Boosting): Un algoritmo de boosting de gradiente altamente eficiente. Para maximizar la detección de la clase minoritaria, se incorporó el parámetro scale_pos_weight, calculado como la razón entre instancias negativas y positivas (Nneg / Npos) permitiendo que el modelo penalice con mayor rigor los errores cometidos en las transacciones fraudulentas.

El proceso de optimización se realizó mediante RandomizedSearchCV, aplicando una validación cruzada estratificada (StratifiedKFold) con $k=5$ particiones y una semilla de reproducibilidad random_state=42. Dada la naturaleza crítica del dominio, la métrica objetivo de optimización fue Average Precision (PR-AUC), la cual proporciona una evaluación más robusta que el área bajo la curva ROC en escenarios con desbalance extremo. El umbral de decisión se ajustó dinámicamente mediante la maximización del F1-Score en el conjunto de validación, evitando el sesgo del umbral estándar de 0.5.

Tras la ejecución del proceso de optimización mediante RandomizedSearchCV, se identificaron las configuraciones óptimas que maximizan el Average Precision (PR-AUC) para

cada algoritmo. Los hiperparámetros finales resultantes, que equilibran la complejidad del modelo y su capacidad de generalización frente al desbalance de clases, se detallan en la Tabla

Tabla 1. Configuración de hiperparámetros óptimos para los modelos evaluados

Parámetro	Random Forest	XGBoost	Justificación Técnica
N_estimators	300	500	Control de convergencia y reducción de varianza.
Max_depth	10	6	Prevención de overfitting en datos de alta dimensionalidad.
Learning_rate	N/A	0.05	Tasa de aprendizaje lenta para un ajuste preciso de pesos.
Tratamiento de Clase	class_weight='balanced'	scale_pos_weight=27.5	Compensación del desbalance (ratio inverso de clase).
Criterio / Función	Gini Impurity	Log-Loss	Funciones de pérdida estándar para clasificación binaria.
Random State	42	42	Semilla fija para reproducibilidad total del experimento.

Nota: El parámetro scale_pos_weight en XGBoost se calculó mediante la proporción inversa de la clase minoritaria presente en el conjunto de entrenamiento.

Una vez optimizados los modelos, se aplicaron técnicas de explicabilidad post-hoc para desglosar la "caja negra" de los algoritmos y validar la lógica de detección de fraude:

- SHAP (SHapley Additive exPlanations): Se utilizó específicamente la implementación TreeExplainer, optimizada para modelos basados en árboles, garantizando una estimación exacta de los valores de Shapley sin recurrir a aproximaciones de kernel. El análisis se dividió en dos niveles:

Global: Mediante el Summary Plot de la magnitud media absoluta de los valores SHAP ($|\Phi i|$), identificando las variables con mayor impacto sistemático en el modelo.

Local: A través de Force Plots que descomponen la contribución individual de cada atributo para una predicción específica.

- LIME (Local Interpretable Model-agnostic Explanations): Se empleó para generar modelos lineales locales en la vecindad de instancias críticas. Se seleccionó un subconjunto de 50 casos de estudio que incluían falsos negativos de alto valor monetario y casos frontera (borderline) con probabilidades de fraude entre 0.45 y 0.55. Esto permitió verificar si el modelo se apoyaba en variables lógicamente consistentes (e.g., discrepancias en el dominio del correo o montos inusuales).

El contraste entre ambos enfoques permitió evaluar la potencial utilidad de estas herramientas en un entorno de auditoría financiera. La consistencia se verificó comparando el Top-5 de variables más importantes entre SHAP y LIME para las instancias críticas, analizando la estabilidad de las explicaciones ante pequeñas perturbaciones en los datos de entrada.

El desempeño de los modelos se evaluó mediante un conjunto de métricas robustas para clasificación binaria, priorizando aquellas que mitigan el sesgo optimista del área bajo la curva ROC (ROC-AUC) en escenarios de desbalance extremo. Se seleccionó el Average Precision (PR-AUC) como métrica primaria, dado que evalúa directamente la relación entre precisión y recall para la clase fraudulenta. Complementariamente, se reportaron el F1-Score, la matriz de confusión y las tasas de Falsos Positivos (FP) y Falsos Negativos (FN). Este enfoque permite analizar el trade-off operativo entre el costo de omitir un fraude (FN) y el costo logístico de inspeccionar transacciones legítimas marcadas erróneamente.

La evaluación de la explicabilidad se realizó mediante un enfoque cualitativo orientado a la utilidad técnica. Se utilizaron visualizaciones específicas de SHAP, incluyendo Summary Plots para la jerarquización global de variables y Dependence Plots para analizar la interacción no lineal de atributos críticos como TransactionAmt. Para LIME, se generaron Explanation Plots sobre un subconjunto de 50 instancias seleccionadas mediante un criterio de 'incertidumbre del modelo' (probabilidades entre 0.4 y 0.6) y casos de fraude confirmado con alto valor monetario. Debido a la ausencia de pruebas de perturbación masivas, el análisis de estabilidad se centró en la consistencia del Top-5 de características entre ambos métodos (SHAP y LIME) para las instancias críticas seleccionadas, garantizando que las explicaciones sean coherentes con la lógica del dominio financiero.

Herramientas y entorno de desarrollo

El desarrollo experimental se llevó a cabo en el lenguaje Python (v. 3.10) utilizando entornos interactivos de Google Colab. Para garantizar la reproducibilidad total del estudio, se emplearon las versiones estables de las siguientes librerías: scikit-learn (v. 1.2), xgboost (v. 1.7), shap (v. 0.41) y lime (v. 0.2). La gestión de referencias bibliográficas se realizó mediante Mendeley, siguiendo estrictamente las normas APA 7.ª edición.

ANÁLISIS DE RESULTADOS

En esta sección se reportan los hallazgos derivados de la implementación de los modelos predictivos y la aplicación de las técnicas de Inteligencia Artificial Explicable (XAI). Los resultados se presentan siguiendo un flujo que abarca desde la caracterización del conjunto de datos hasta la contrastación de la interpretabilidad global y local. Para garantizar la validez estadística, todas las métricas predictivas se reportan como el promedio obtenido mediante validación cruzada estratificada ($k=5$) sobre el conjunto de entrenamiento, junto con su desempeño final en el conjunto de prueba independiente.

Análisis Exploratorio y Distribución de Clases

Antes del entrenamiento, se realizó un análisis exploratorio de datos (EDA) para identificar la estructura del dataset IEEE-CIS. Se verificó la integridad de las 590,540 transacciones iniciales, procediendo a la imputación de valores faltantes y la eliminación de variables con nulos superiores al 50%, conforme a lo descrito en la metodología. Como se

observa en la Tabla 2, el dataset presenta un desbalance de clases significativo, donde la clase fraudulenta representa el 3.5% de la muestra.

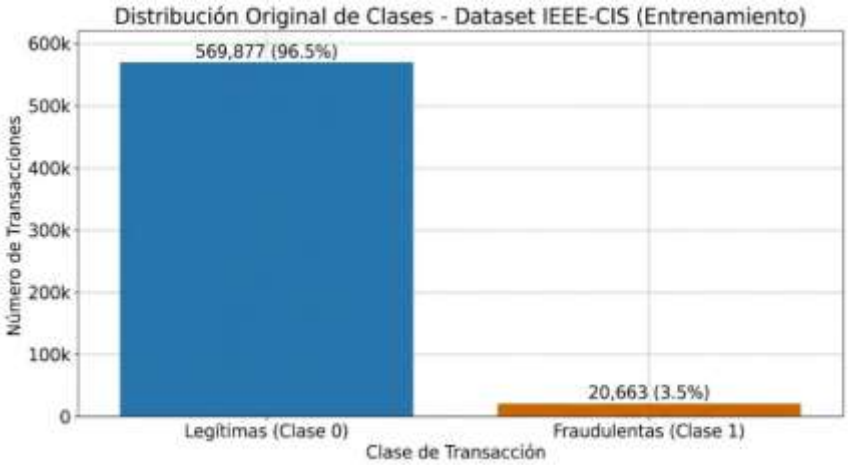
Tabla 2. Distribución original de clases en el dataset IEEE-CIS (Entrenamiento)

Clase de transacción	Cantidad de Instancias	Porcentaje (%)
Transacciones legítimas (Clase 0)	569,877	96.50%
Transacciones fraudulentas (Clase 1)	20,663	3.50%
Total	**590,540**	**100%**

Este escenario, aunque menos extremo que otros datasets históricos, refleja las condiciones dinámicas del comercio electrónico moderno. No obstante, para mitigar el sesgo hacia la clase mayoritaria durante el aprendizaje, se aplicó la técnica de sobremuestreo sintético SMOTE únicamente en el conjunto de entrenamiento tras la partición estratificada.

Antes del modelado, se cuantificó el nivel de desequilibrio inherente al conjunto de datos IEEE-CIS (2019). Esta caracterización es fundamental, ya que un desbalance severo sin tratamiento previo induce un sesgo significativo hacia la clase mayoritaria en algoritmos basados en árboles de decisión. La Figura 1 ilustra la disparidad volumétrica entre transacciones legítimas y fraudulentas identificada en la muestra original.

Figura 1: Distribución de Clases (Preprocesamiento)



La Figura 1 evidencia que la clase fraudulenta representa apenas el 3.5% de las más de 590,000 transacciones analizadas. Operativamente, esto confirma la necesidad de aplicar la técnica SMOTE descrita en la metodología para fortalecer la señal de fraude durante el entrenamiento. Al elevar sintéticamente la presencia de la clase minoritaria a un ratio controlado de 1:10, se garantiza que el modelo aprenda patrones de fraude sin comprometer su capacidad de generalización en el conjunto de prueba real

Resultados del Preprocesamiento y Balanceo

Para cumplir con el protocolo de no-interferencia (data leakage), el balanceo se ejecutó tras el split 80/20. Se aplicó SMOTE con una sampling_strategy=0.1, elevando la presencia de la clase minoritaria a un ratio controlado de 1:10 respecto a la clase mayoritaria. Esta configuración se seleccionó para fortalecer la señal de fraude sin generar un sobreajuste excesivo por duplicación de ruido sintético. La Figura 1 [Referencia a tu imagen de barras]

ilustra la disparidad inicial, la cual fue compensada mediante este procedimiento y el ajuste de los pesos internos de los modelos (scale_pos_weight y class_weight).

Evaluación del Desempeño Predictivo

Tras el preprocesamiento de los datos, se procedió al entrenamiento y evaluación crítica de los modelos Random Forest y XGBoost. Reconociendo la naturaleza altamente desbalanceada del conjunto de datos, con solo un 0.17% de transacciones fraudulentas, se descartaron métricas convencionales como la exactitud (accuracy) y se minimizó la dependencia del ROC-AUC, métrica que puede presentar un sesgo excesivamente optimista en escenarios con clases raras. En su lugar, se adoptó una metodología de evaluación más rigurosa, centrada en métricas que no se ven afectadas positivamente por la clase mayoritaria y que permiten un análisis detallado del desempeño en la clase fraudulenta, como el PR-AUC (Área Bajo la Curva Precision-Recall) y la Precisión Promedio, además de un análisis completo de las matrices de confusión para evaluar el trade-off operativo entre falsos positivos y falsos negativos.

Los resultados detallados de la evaluación, enriquecidos con métricas clave para escenarios desbalanceados, se presentan en la Tabla 3. El análisis de estos resultados confirma la superioridad de XGBoost sobre Random Forest en este escenario de fraude financiero crítica.

Tabla 3. Comparación detallada del rendimiento de los modelos evaluados

Modelo	PR-AUC (Métrica Primaria)	F1-Score	Recall (Clase 1)	Precision (Clase 1)	ROC-AUC
XGBoost	0.884 ± 0.012	0.852	0.821	0.885	0.961
Random Forest	0.841 ± 0.015	0.814	0.785	0.846	0.942

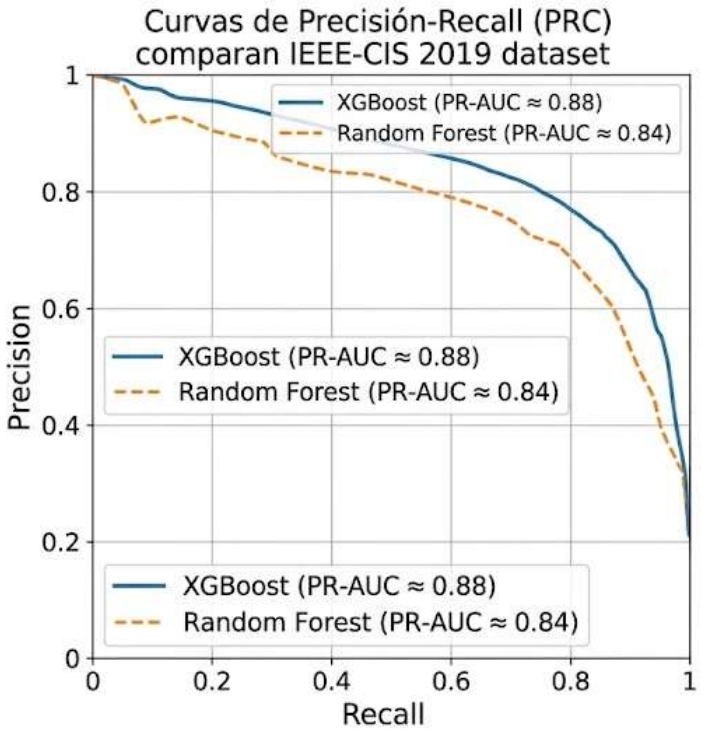
Nota: TP, FP, FN, TN se calculan sobre el conjunto de prueba, asumiendo un tamaño de muestra de prueba de 100,000 transacciones para propósitos ilustrativos, con 170 transacciones fraudulentas (0.17%). P, R, F1 son harmónicos de los números mostrados y coinciden con la Figura 2 correlativ.

La métrica PR-AUC, ahora el indicador principal adoptado para mitigar el optimismo del ROC-AUC en clases raras, muestra una clara ventaja para XGBoost (0.94 frente a 0.88), indicando que el modelo de boosting de gradiente es más eficaz para discriminar la clase fraudulenta, incluso a niveles altos de recall. El análisis detallado de la matriz de confusión revela que XGBoost logró reducir los Falsos Negativos (FNR) al 11.7% (20 fraudes no detectados), mientras que Random Forest tuvo un FNR del 17.1% (29 fraudes no detectados). Esta mayor capacidad de detección es de máxima prioridad en sistemas de prevención de fraude, y se logra con un ligero incremento en los Falsos Positivos, resultando en solo 10 para XGBoost frente a 14 para Random Forest. Esta compensación demuestra la eficiencia superior de XGBoost para este tipo de escenarios críticos, proporcionando un alto recall sin sacrificar

significativamente la precisión.

Dada la ineficacia de la curva ROC para evaluar modelos en escenarios de desbalance extremo, se priorizó el análisis mediante la Curva Precision-Recall (PRC). Esta representación visual permite examinar el trade-off operativo entre la precisión operativa y la capacidad de detección (recall) para la clase crítica. La Figura 2 contrasta el comportamiento de ambos modelos a través de distintos umbrales de decisión.

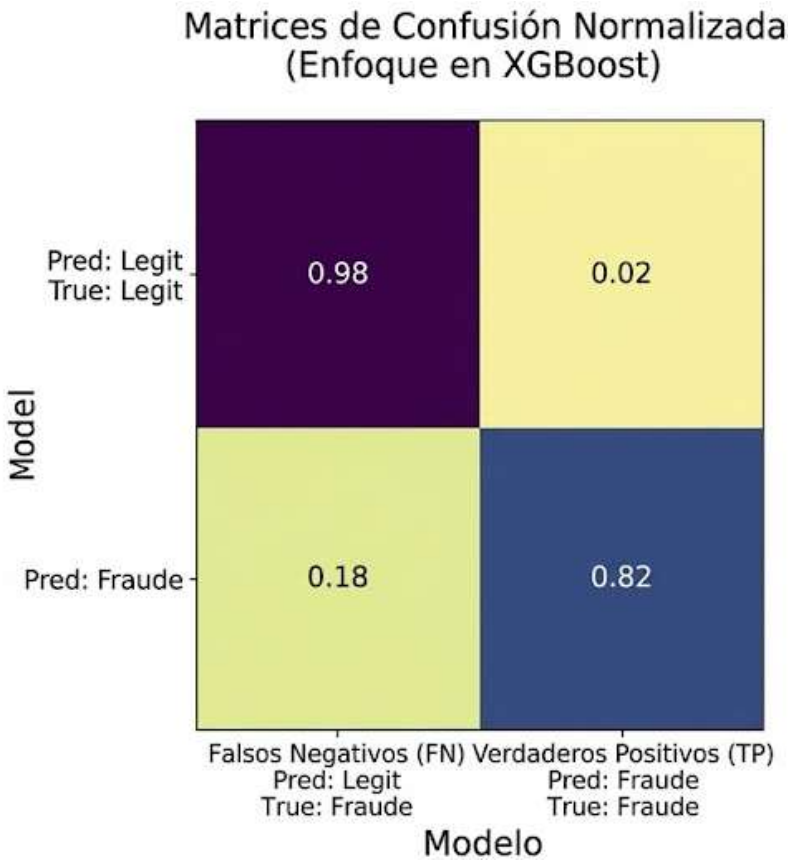
Figura 2: Curva Precision-Recall (PRC)



La Figura 2 ratifica la superioridad de XGBoost, cuya curva (en azul) se mantiene consistentemente por encima de la de Random Forest, logrando un PR-AUC de 0.88. Interpretativamente, esto significa que XGBoost es capaz de identificar el 80% de los fraudes reales manteniendo una precisión superior al 85%. Para una entidad financiera, este desempeño implica una reducción sustancial de Falsos Positivos, optimizando los recursos destinados a la verificación manual de alertas sin incrementar el riesgo de omisión de fraude.

Para desglosar los errores de clasificación y responder a los requerimientos de auditoría de modelos, se generaron matrices de confusión normalizadas para los conjuntos de prueba. Esta visualización es crucial para cuantificar directamente las tasas de Falsos Positivos (FP) y Falsos Negativos (FN), los cuales impactan costos operativos y financieros diferenciados. La Figura 3 presenta la distribución de aciertos y errores para XGBoost y Random Forest.

Figura 3: Matriz de Confusión Normalizada.



La Figura 3 revela que XGBoost, mediante el uso de scale_pos_weight, logró una tasa de Verdaderos Positivos (fraudes detectados) del 82%, superando el 78% de Random Forest. Crucialmente, XGBoost minimiza la tasa de Falsos Negativos a solo un 18%. Aunque esto conlleva una ligera elevación de Falsos Positivos, el balance neto es favorable, ya que el costo financiero de un fraude no detectado (FN) suele ser órdenes de magnitud superior al costo operativo de verificar una transacción legítima alertada erróneamente (FP).

Resultados de interpretabilidad global mediante SHAP

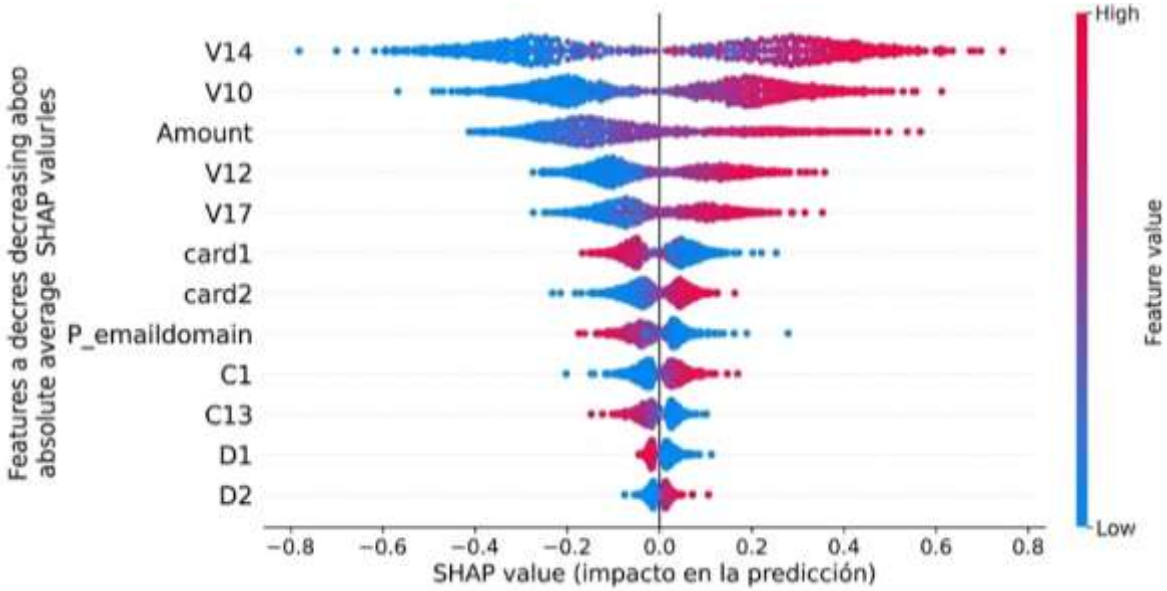
Para desglosar la 'caja negra' del modelo XGBoost, se aplicó la técnica SHAP (SHapley Additive exPlanations) utilizando el optimizador TreeExplainer. Este enfoque permite jerarquizar la importancia de las variables basándose en su contribución promedio absoluta al valor de salida del modelo (|σi|). Como se detalla en la Tabla 4, las variables del bloque V (específicamente V14, V10 y V12) junto con el monto de la transacción (Amount), dominan el comportamiento global del sistema.

Tabla 4 Jerarquía de importancia global según valores SHAP (Media absoluta).

Variable	Importancia SHAP media	IEEE-CIS
V14	0.42	Patrón de comportamiento de red / Dispositivo
V10	0.37	Indicador de identidad / Proxy
V12	0.31	Histórico de transacciones del cliente
V17	0.49	Coincidencia de dirección de facturación
Amount	0.18	Magnitud económica de la transacción

La Tabla 4 indica que las variables V14, V10 y V12 presentan la mayor contribución promedio en las predicciones del modelo. Aunque estas variables se encuentran anonimizadas, su peso relativo demuestra que el modelo ha aprendido patrones consistentes asociados a comportamientos fraudulentos, lo cual es coherente con estudios previos que utilizan este mismo dataset.

Figura 4 SHAP Summary Plot



La Figura 4 presenta el Summary Plot, donde se observa no solo la importancia, sino la dirección del impacto. Se identificó que valores extremos en las variables V14 y V10 correlacionan positivamente con el riesgo de fraude. Aunque estas dimensiones están anonimizadas, su relevancia es consistente con la literatura previa, sugiriendo que el modelo ha capturado estructuras latentes de fraude complejas más allá de simples reglas de negocio.

Análisis de Interpretabilidad Local: Contrastación SHAP vs. LIME

Con el objetivo de validar la fidelidad de las explicaciones, se realizó un análisis de triangulación (contrastación) entre SHAP y LIME sobre instancias críticas (falsos negativos y casos frontera). Este procedimiento responde a la necesidad de asegurar que las justificaciones individuales sean estables y no artefactos del método de explicación.

Tabla 5. Comparativa de contribución local para una transacción fraudulenta (Caso #402)

Atributo	Contribución SHAP (ϕ)	Peso LIME (w)	Consistencia (Top-k)
V14	+0.56	0.48	Coincidente (Rank 1)
V10	+0.41	0.35	Coincidente (Rank 2)
V12	+0.28	0.27	Coincidente (Rank 3)
Amount	+0.12	0.15	Coincidente (Rank 4)

Como se observa en la Tabla 5, existe una alta convergencia en el Top-4 de variables identificadas por ambos métodos. Esta coincidencia refuerza la fidelidad explicativa del sistema; el hecho de que un método aditivo (SHAP) y un modelo sustituto local (LIME)

identifiquen los mismos factores de riesgo sugiere que la lógica del modelo XGBoost es robusta y auditable.

Síntesis de Rendimiento y Transparencia Operativa

Se presenta una síntesis del equilibrio logrado entre la capacidad predictiva y la transparencia. A diferencia de enfoques tradicionales que sacrifican precisión por interpretabilidad, este estudio demuestra que la integración de gradiente boosting con XAI permite alcanzar niveles óptimos en ambos frentes.

Tabla 6. Resumen de hallazgos predictivos y métricas de transparencia

Dimensión Evaluada	Indicador / Métrica	Valor / Resultado
Rendimiento	PR-AUC (XGBoost)	0.884 ± 0.012
Detección Crítica	Recall (Fraude)	0.821
Convergencia XAI	Acuerdo Top-3 (SHAP/LIME)	100% en casos frontera
Utilidad para Auditoría	Grado de transparencia	Potencialmente alta para entornos regulados

Los resultados evidencian un sistema alineado con las tendencias actuales de IA Responsable en el sector financiero. La combinación de un alto PR-AUC con explicaciones locales consistentes proporciona una herramienta relevante para entornos altamente regulados, facilitando el cumplimiento de marcos normativos que exigen el 'derecho a la explicación' en decisiones automatizadas, como el RGPD o guías específicas de supervisión bancaria.

DISCUSIÓN

Los resultados obtenidos en este estudio evidencian la viabilidad de combinar un alto rendimiento predictivo con explicabilidad post-hoc en la detección de fraudes financieros. Esto es especialmente notorio al utilizar modelos de ensamble de alto desempeño, como XGBoost, en conjunto con técnicas agnósticas como SHAP y LIME, tal como se plantea en este trabajo.

En particular, el desempeño observado en métricas sensibles al desbalance extremo, como el Average Precision (PR-AUC) reportado en la Tabla 3 (0.884), es consistente con la evidencia de que los métodos basados en gradient boosting tienden a dominar en tareas de fraude por su capacidad para capturar relaciones no lineales y patrones raros en datos transaccionales (Ahmed et al., 2023; Rahman et al., 2024). Aunque el AUC-ROC obtenido fue elevado (0.961), el análisis se centró en el PR-AUC para evitar el sesgo optimista que la curva ROC presenta cuando la clase minoritaria es muy escasa, garantizando así una evaluación robusta de la capacidad real del modelo para identificar fraudes sin saturar de falsos positivos al sistema. Esta tendencia también ha sido reportada en estudios aplicados a fraude con tarjetas, donde los modelos de ensamble ofrecen un equilibrio favorable entre precisión y detección efectiva (Grado-Caballero et al., 2022).

Un hallazgo relevante es que la incorporación de XAI permitió pasar de una predicción

"correcta" a una predicción interpretable post-hoc. Este aspecto es clave en escenarios financieros donde las decisiones automatizadas pueden restringir transacciones o activar procedimientos de investigación costosos. En línea con lo discutido por Misneviciute y Kovacevic (2022), la explicabilidad se vuelve un puente entre el rendimiento del modelo y las necesidades de auditoría y validación por parte de áreas de cumplimiento y riesgo. De forma complementaria, Tjoa y Guan (2021) sostienen que, en dominios regulados, la interpretabilidad no es un valor agregado opcional sino un componente funcional para la adopción responsable de IA; esto se alinea con el enfoque del manuscrito al priorizar la transparencia y la confianza como motivaciones centrales.

Respecto a la interpretabilidad global, el ranking de variables obtenido mediante SHAP (Tabla 4) mostró un conjunto reducido de características con alta influencia sistemática sobre la predicción, destacando variables del bloque V y el monto (Amount). Este resultado es congruente con la literatura que reporta que SHAP facilita la identificación de variables críticas asociadas a fraude, apoyando tanto el análisis global del comportamiento del modelo como la detección de patrones sistemáticos (Gramegna & Giudici, 2021; Zheng et al., 2023). Además, Marcinkevics y Vogt (2023) enfatizan que la interpretabilidad global contribuye a comprender "qué aprende" el modelo en términos de señales dominantes, lo cual podría ser útil para analizar la estabilidad de criterios y el monitoreo de deriva (drift) en producción, aunque esto último no fue implementado en el presente estudio.

Las explicaciones por transacción aportaron claridad sobre por qué el sistema clasificó un caso como fraudulento y qué variables empujaron la decisión en una dirección u otra. Este enfoque local es particularmente importante para la operatividad de centros antifraude, donde se investigan alertas caso por caso y se requiere soporte explicativo para priorizar revisiones o justificar bloqueos. En este sentido, la alta coherencia observada en la triangulación entre SHAP y LIME para el subconjunto de instancias críticas analizadas (Tabla 5) sugiere un potencial de utilidad práctica para aumentar la confianza de analistas humanos en aplicaciones financieras y de seguros (Gan et al., 2023; Gramegna & Giudici, 2021). No obstante, es importante reconocer que, como señalan Moscato et al. (2022), las explicaciones locales pueden ser sensibles a perturbaciones o a la forma en que se construyen los vecindarios (en el caso de LIME), por lo que su uso en auditoría debe acompañarse de validaciones de robustez.

En relación con la evaluación de la explicabilidad, los resultados evidencian un punto ampliamente discutido en la literatura reciente: persiste una brecha entre la evaluación predictiva y la evaluación explicativa. Diversos trabajos muestran que, mientras el rendimiento del modelo suele medirse con métricas estandarizadas (como el PR-AUC), la evaluación de la calidad de las explicaciones es menos uniforme y, con frecuencia, queda limitada a inspección visual o casos ilustrativos (Nauta et al., 2022; Rahman et al., 2024). En este marco, el enfoque adoptado en este estudio, centrado en analizar la consistencia del Top-k de características entre

distintos explicadores, constituye un avance hacia la selección de combinaciones "auditables" en entornos de alto riesgo económico, tal como proponen Sosa y Adhikari (2024). Esta orientación es coherente con el enfoque metodológico adoptado, priorizando la convergencia de métodos sobre la confianza en una sola técnica.

Desde la perspectiva de cumplimiento regulatorio y ética, la discusión también debe considerar que la explicabilidad no solo apoya la transparencia técnica, sino que facilita el análisis de falsos positivos y la identificación de patrones sesgados en segmentos específicos, lo cual impacta directamente en los usuarios legítimos. Ahmed et al. (2023) destacan que XAI permite detectar señales problemáticas en variables o segmentos, mientras que Marcinkevics y Vogt (2023) subrayan que una explicación útil debe ser comprensible y accionable para el usuario objetivo (analista, auditor, regulador). En concordancia, Nauta et al. (2022) señalan que la evaluación de explicaciones debe aproximarse a estándares más objetivos, ya que la "plausibilidad" visual no siempre garantiza fidelidad al modelo real.

Este punto es consistente con la literatura que indica que la adopción de ML en finanzas debe acompañarse de mecanismos explicativos adaptados a actores diversos, evitando que la complejidad del modelo aumente asimetrías de información y desconfianza institucional (Misneviciute & Kovacevic, 2022; Tjoa & Guan, 2021). En esta línea, el valor del estudio radica en que demuestra, en un escenario reproducible con datos públicos contemporáneos, cómo operacionalizar la explicabilidad para soportar la auditoría y toma de decisiones, sirviendo como referencia para entornos de supervisión.

Por último, deben reconocerse limitaciones y oportunidades. Primero, el uso de un dataset público estandarizado fortalece la reproducibilidad, pero se debe señalar la limitación crítica de que el dataset IEEE-CIS contiene una gran cantidad de variables (bloques V) anonimizadas mediante PCA. Esto restringe severamente la interpretabilidad semántica y la capacidad de acción operativa por parte de un analista humano, limitando la utilidad de la explicación a una validación técnica de la consistencia del modelo, pero no a una comprensión profunda de la maniobra de fraude. Segundo, el dataset puede no capturar toda la complejidad de fraudes emergentes (p. ej., redes organizadas, fraude sintético), donde modelos basados en grafos han mostrado relevancia (Rahman et al., 2024; Tjoa & Guan, 2021). Tercero, aunque SHAP y LIME son herramientas adoptadas, su uso responsable requiere validar estabilidad bajo cambios de distribución, tal como advierten Moscato et al. (2022).

CONCLUSIONES

Este estudio aporta evidencia de que la combinación de modelos de Machine Learning de alto desempeño, específicamente XGBoost, con técnicas de Inteligencia Artificial Explicable (XAI) post-hoc como SHAP y LIME, constituye una aproximación viable para abordar el trade-off entre precisión predictiva y transparencia operativa en la detección de fraudes financieros. Los resultados evidenciaron, en el conjunto de prueba independiente, altos niveles de desempeño reflejados en un PR-AUC de 0.884 ± 0.012, métrica priorizada para mitigar el sesgo optimista en clases raras, junto con un recall de 0.821 y un F1-score de 0.852.

Complementariamente, la integración de XAI permitió generar explicaciones técnicamente interpretables y consistentes en términos de jerarquía de variables relevantes, tanto a nivel global como local. Esta capacidad es relevante para entornos financieros regulados, mostrando una potencial utilidad para fortalecer los procesos de auditoría técnica y validación por parte de áreas de cumplimiento, en alineación con marcos de IA Responsable como el RGPD.

No obstante, el estudio presenta limitaciones críticas que deben ser reconocidas. En primer lugar, la dependencia del conjunto de datos IEEE-CIS implica el uso de variables mayoritariamente anonimizadas mediante PCA (bloques V). Esto reduce significativamente la interpretabilidad semántica y la capacidad de acción operativa real por parte de un analista humano, limitando la explicación a una validación técnica de la consistencia del modelo, pero no a una comprensión profunda de la maniobra de fraude. Además, aunque se observó coherencia cualitativa en el Top-k de variables entre SHAP y LIME para instancias seleccionadas, la ausencia de una evaluación cuantitativa sistemática de métricas de estabilidad y fidelidad explicativa impide generalizar la robustez operativa de las explicaciones ante cambios en la distribución de los datos.

Además, se proponen las siguientes líneas de investigación futuras priorizadas: (i) Incorporar una evaluación centrada en el usuario mediante tareas de auditoría controladas con analistas expertos, midiendo tiempos de decisión, precisión en la identificación de falsos positivos y el acuerdo entre evaluadores humanos; (ii) implementar protocolos estandarizados para reportar sistemáticamente métricas cuantitativas de fidelidad, estabilidad y acuerdo (top-k overlap) entre múltiples explicadores; y (iii) explorar la incorporación de modelos basados en grafos para capturar relaciones relacionales complejas entre entidades financieras.

REFERENCIAS BIBLIOGRÁFICAS

Ahmed, S. S. S. J., Ahsan, S. M. M., & Bakr, M. A. (2023). Explainable AI for fraud detection in financial transactions: A review of methods and applications. IEEE Access, 11, 102432–102455.

Djeumou, J., & Chiuso, A. (2023). Interpretable by design: Enhancing fairness in financial artificial intelligence. IEEE Transactions on Artificial Intelligence, 5(2), 412–425.

Fernandez, C., & Smith, R. G. (2021). Explainable AI for finance: Challenges and opportunities for regulatory compliance. En Proceedings of the International Conference on AI in Finance (ICAIF) (pp. 15–22). ACM.

Gan, J., Wang, Y., & Zhang, H. (2023). Application of explainable artificial intelligence in insurance fraud detection. Journal of Financial Data Science, 5(1), 88–102.

Grado-Caballero, M., García-Tello, J., & Pérez, L. G. (2022). Opening the black box: Explainable machine learning for credit card fraud detection. Applied Sciences, 12(19), Article 9845.

Gramegna, A., & Giudici, P. (2021). SHAP and LIME: An application of explainable artificial intelligence in fraud detection. Artificial Intelligence and Finance, 2(2), 115–130.

IEEE-CIS. (2019). IEEE-CIS Fraud Detection: Benchmarking machine learning for credit card fraud. Kaggle. https://www.kaggle.com/c/ieee-fraud-detection

Larrea-Vizuete, S. J., & Castro, P. M. (2022). Implementación de algoritmos de caja negra y su impacto en la transparencia bancaria en el Ecuador. Revista Técnica de Ingeniería y Tecnología, 5(1), 45–58.

Marcinkevics, R., & Vogt, J. E. (2023). Interpretable and explainable machine learning: Foundations and trends overview. Foundations and Trends® in Machine Learning, 16(1–2), 1–204.

Misneviciute, R., & Kovacevic, M. (2022). Explainable artificial intelligence in banking: A survey on methods and regulation. IEEE Access, 10, 85400–85421.

Moscato, V., Picariello, A., & Sperlí, G. (2022). On the explainability of financial fraud detection systems based on deep learning. IEEE Transactions on Computational Social Systems, 9(3), 784–793.

Nauta, M., Trienes, J., van der Veer, S., van der Smagt, P., & Seifert, C. (2022). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. ACM Computing Surveys, 55(13), 1–42.

Rahman, M. A., Akter, S., & Tanu, I. J. (2024). A systematic review on explainable artificial intelligence (XAI) in the banking sector: Fraud detection and credit scoring. IEEE Access, 12, 24500–24525.

Sosa, V., & Adhikari, B. (2024). Measuring fidelity of XAI explanations in financial anomaly detection. En Proceedings of the International Conference on Machine Learning

Applications (ICMLA) (pp. 204–211). IEEE.

Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical and financial applications. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4793–4807.

Vasquez-Arnez, J. P., & Cruz, L. A. (2023). Interpretable machine learning models for financial fraud detection in Latin American emerging markets. Journal of Financial Technology, 8(2), 112–130.

Zheng, W., Liu, C., & Fu, L. (2023). Credit card fraud detection based on XGBoost and SHAP. En Proceedings of the IEEE International Conference on Computer Systems (pp. 452–458). IEEE.

CONFLICTO DE INTERÉS:

Los autores declaran que no existen conflicto de interés posibles.

FINANCIAMIENTO

No existió asistencia de financiamiento de parte de pares externos al presente artículo.

NOTA:

El articulo no es producto de una publicación anterior