Ciencia Latina Revista Científica Multidisciplinar, Ciudad de México, México.
ISSN 2707-2207 / ISSN 2707-2215 (en línea), marzo-abril 2026,
Volumen 10, Número 2.

https://doi.org/10.37811/cl_rcm.v10i2

EVALUACIÓN DE TÉCNICAS DE EXPLICABILIDAD EN MODELOS DE APRENDIZAJE AUTOMÁTICO APLICADOS A DECISIONES CRÍTICAS EN INSTITUCIONES

EVALUATION OF EXPLAINABILITY TECHNIQUES IN MACHINE LEARNING MODELS APPLIED TO CRITICAL DECISION-MAKING IN INSTITUTIONAL CONTEXTS

Maria Teodolinda Ortega Ovalle
Universidad de Panamá

DOI: https://doi.org/10.37811/cl_rcm.v10i2.23172

Evaluación de Técnicas de Explicabilidad en Modelos de Aprendizaje Automático Aplicados a Decisiones Críticas en Instituciones

Maria Teodolinda Ortega Ovalle[1]
maria.ortegao@up.ac.pa
https://orcid.org/0009-0000-3629-9751
Universidad de Panamá
Facultad de Informática, Electrónica y Comunicación
Departamento de Informática

RESUMEN

El uso creciente de modelos de aprendizaje automático en instituciones públicas y privadas ha impulsado la automatización de procesos y la mejora en la toma de decisiones, pero también ha generado preocupaciones relacionadas con la transparencia y la capacidad de auditar decisiones críticas. En contextos donde las predicciones afectan directamente a personas o recursos institucionales, la falta de explicabilidad en modelos complejos puede limitar la confianza, dificultar la detección de sesgos y comprometer la validez de los resultados. El objetivo de este trabajo es evaluar diversas técnicas de explicabilidad aplicadas a modelos de aprendizaje automático utilizados en decisiones institucionales de alto impacto. La metodología incluye el entrenamiento de modelos predictivos sobre un conjunto de datos institucional y la aplicación sistemática de técnicas como SHAP, LIME, modelos sustitutos e Integrated Gradients, evaluando métricas de fidelidad, estabilidad, tiempo de cómputo y utilidad práctica para la auditoría algorítmica. Los resultados muestran diferencias significativas entre las técnicas analizadas, evidenciando que algunas ofrecen explicaciones más consistentes y útiles para la interpretación de decisiones críticas. Los hallazgos sugieren que la integración de técnicas de explicabilidad puede fortalecer la transparencia, mejorar la confianza institucional y apoyar la toma de decisiones informadas, en concordancia con los lineamientos éticos y metodológicos recomendados por APA 7

Palabras clave: explicabilidad, interpretabilidad, aprendizaje automático, auditoría algorítmica, transparencia institucional

[1] Autor principal
Correspondencia: maria.ortegao@up.ac.pa

Evaluation of Explainability Techniques in Machine Learning Models Applied to Critical Decision-Making in Institutional Contexts

ABSTRACT

The increasing adoption of machine learning models in public and private institutions has enhanced process automation and decision-making capabilities, yet it has also raised concerns regarding transparency and the auditability of critical decisions. In contexts where predictions directly affect individuals or institutional resources, the lack of explainability in complex models can undermine trust, hinder the detection of biases, and compromise the validity of outcomes. The objective of this study is to evaluate several explainability techniques applied to machine learning models used in high-impact institutional decision-making. The methodology involves training predictive models on an institutional dataset and systematically applying techniques such as SHAP, LIME, surrogate models, and Integrated Gradients, assessing metrics including fidelity, stability, computational cost, and practical usefulness for algorithmic auditing. The results reveal significant differences among the techniques, showing that some provide more consistent and actionable explanations for interpreting critical decisions. The findings suggest that integrating explainability techniques can enhance transparency, strengthen institutional trust, and support informed decision-making, aligning with ethical and methodological guidelines consistent with APA 7 standards

***Keywords*:** explainability, interpretability, machine learning, algorithmic auditing, institutional transparency

Artículo recibido 20 febrero 2026

Aceptado para publicación: 29 marzo 2026

INTRODUCCIÓN

El uso de modelos de aprendizaje automático en instituciones públicas y privadas se ha convertido en una práctica cada vez más extendida para apoyar procesos de evaluación, clasificación, asignación de recursos y toma de decisiones estratégicas. Estos modelos permiten procesar grandes volúmenes de información y generar predicciones con altos niveles de precisión, lo que los convierte en herramientas valiosas para mejorar la eficiencia institucional. Sin embargo, la creciente complejidad de los algoritmos empleados, especialmente aquellos considerados de caja negra, plantea un problema central: la dificultad para comprender, justificar y auditar las decisiones generadas por estos sistemas. Esta falta de transparencia constituye un vacío crítico en el conocimiento aplicado, ya que limita la capacidad de las instituciones para garantizar decisiones confiables, éticas y verificables.

El problema de investigación se centra en la ausencia de mecanismos claros que permitan explicar de manera comprensible cómo los modelos de aprendizaje automático producen sus resultados, especialmente en contextos donde las decisiones tienen consecuencias directas sobre personas, procesos o recursos institucionales. La relevancia de abordar este tema radica en que la explicabilidad no solo fortalece la confianza en los sistemas automatizados, sino que también facilita la detección de sesgos, errores y patrones no deseados que podrían afectar la equidad y la validez de las decisiones. En este sentido, la explicabilidad se convierte en un componente esencial para la gobernanza algorítmica y la responsabilidad institucional.

El estudio se sustenta en teorías y enfoques provenientes de la ciencia de datos, la inteligencia artificial explicable y la auditoría algorítmica. Conceptos como modelos de caja negra, interpretabilidad local y global, fidelidad de las explicaciones y estabilidad de los métodos de interpretación constituyen el marco conceptual que orienta el análisis. Autores como Ribeiro, Lundberg, Doshi-Velez y Molnar han desarrollado técnicas y fundamentos teóricos que permiten comprender cómo se generan las explicaciones y qué criterios deben considerarse para evaluar su calidad. Estas teorías proporcionan las categorías analíticas necesarias para comparar diferentes métodos de explicabilidad y valorar su utilidad en entornos institucionales.

Diversos estudios previos han explorado la aplicación de técnicas de explicabilidad en sectores como la salud, las finanzas y la administración pública, destacando la importancia de contar con herramientas

que permitan interpretar decisiones automatizadas. Sin embargo, la literatura muestra que aún existe una brecha significativa en la evaluación comparativa de estas técnicas dentro de contextos institucionales específicos, donde las decisiones críticas requieren altos niveles de transparencia y justificación. Este trabajo aporta a dichos antecedentes mediante un análisis sistemático de varias técnicas de explicabilidad aplicadas a modelos predictivos utilizados en instituciones, con el fin de identificar sus fortalezas, limitaciones y potencial de uso real.

La investigación se desarrolla en un contexto institucional donde la toma de decisiones automatizadas adquiere relevancia creciente debido a la necesidad de gestionar información compleja y garantizar procesos eficientes. Este contexto incluye consideraciones éticas, normativas y operativas que influyen en la adopción de modelos explicables y en la evaluación de su impacto. Finalmente, el estudio plantea como objetivo general evaluar la utilidad y desempeño de diversas técnicas de explicabilidad aplicadas a modelos de aprendizaje automático utilizados en decisiones críticas institucionales. Los objetivos específicos se orientan a comparar métricas de fidelidad y estabilidad, analizar la utilidad práctica de las explicaciones y determinar el aporte de cada técnica a la transparencia y la auditoría algorítmica.

METODOLOGÍA

La metodología utilizada en este estudio se desarrolla bajo un enfoque cuantitativo, dado que se analizan datos estructurados y se evalúan métricas numéricas asociadas al desempeño de modelos de aprendizaje automático y de las técnicas de explicabilidad aplicadas. El tipo de investigación es de carácter descriptivo y explicativo, ya que busca describir el comportamiento de diferentes técnicas de explicabilidad y explicar su utilidad en la interpretación de decisiones críticas dentro de instituciones. Asimismo, el estudio posee un componente aplicativo, pues se orienta a la evaluación práctica de métodos que pueden ser implementados en contextos institucionales reales.

El diseño de investigación es experimental, debido a que se manipulan modelos predictivos y se aplican técnicas de explicabilidad bajo condiciones controladas para observar sus efectos en métricas como fidelidad, estabilidad y tiempo de cómputo. El estudio es transversal, ya que el análisis se realiza en un único momento temporal utilizando un conjunto de datos institucional previamente recopilado. La población de estudio está constituida por registros institucionales relacionados con procesos de evaluación y toma de decisiones.

A partir de esta población se selecciona una muestra no probabilística por conveniencia, compuesta por los datos disponibles y validados para el entrenamiento y evaluación de los modelos.

La recolección de datos se basa en la revisión documental de registros institucionales anonimizados, los cuales incluyen variables relevantes para la construcción de modelos predictivos. El procesamiento de datos se realiza mediante herramientas de análisis estadístico y bibliotecas de aprendizaje automático. Los materiales utilizados incluyen un conjunto de datos estructurados, software especializado para el entrenamiento de modelos y librerías para la implementación de técnicas de explicabilidad como SHAP, LIME, modelos sustitutos e Integrated Gradients. No se emplean instrumentos de recolección como encuestas o entrevistas, dado que el estudio se basa exclusivamente en datos institucionales existentes. Las consideraciones éticas del estudio se fundamentan en el uso responsable de datos institucionales, garantizando la anonimización de la información y el cumplimiento de principios de confidencialidad y protección de datos. Los criterios de inclusión consideran únicamente registros completos y consistentes, mientras que se excluyen aquellos con información faltante o inconsistencias que puedan afectar la validez del análisis. Entre las limitaciones del estudio se reconoce la dependencia de un único conjunto de datos institucional, lo que puede restringir la generalización de los resultados a otros contextos. Sin embargo, la metodología empleada permite replicar el estudio en diferentes instituciones o dominios, favoreciendo la comparabilidad y la evaluación futura de técnicas de explicabilidad en escenarios diversos.

MARCO TEÓRICO

El desarrollo de modelos de aprendizaje automático en instituciones se enmarca en la evolución de la ciencia de datos como disciplina orientada al análisis de grandes volúmenes de información para apoyar procesos de decisión. Estos modelos permiten identificar patrones, predecir comportamientos y optimizar recursos, pero su creciente complejidad ha generado la necesidad de comprender cómo producen sus resultados. En este contexto surge el campo de la inteligencia artificial explicable, cuyo propósito es ofrecer mecanismos que permitan interpretar y justificar las decisiones generadas por algoritmos, especialmente aquellos considerados de caja negra.

El concepto de modelo de caja negra se refiere a sistemas cuyo funcionamiento interno no es fácilmente interpretable por los usuarios, ya sea por su estructura matemática, su nivel de abstracción o la cantidad

de parámetros involucrados. Modelos como redes neuronales profundas, máquinas de soporte vectorial y algoritmos de ensamble suelen presentar este tipo de opacidad. La falta de interpretabilidad plantea desafíos éticos y operativos, ya que dificulta la auditoría, la detección de sesgos y la validación de decisiones en entornos institucionales donde la transparencia es un requisito fundamental.

La inteligencia artificial explicable se sustenta en teorías y enfoques que buscan generar explicaciones comprensibles para usuarios técnicos y no técnicos. Entre los aportes más relevantes se encuentran los trabajos de Ribeiro, que propone LIME como un método para generar explicaciones locales mediante modelos simples que aproximan el comportamiento del modelo original; Lundberg y Lee, quienes desarrollan SHAP basado en valores de Shapley provenientes de la teoría de juegos; y Molnar, quien sistematiza los métodos existentes y propone categorías analíticas para su estudio. Estas teorías proporcionan las bases conceptuales necesarias para comparar diferentes métodos de explicabilidad y valorar su utilidad en entornos institucionales.

A continuación, se presenta una tabla comparativa que sintetiza las características principales de las técnicas de explicabilidad utilizadas en este estudio, lo que permite comprender sus diferencias conceptuales y su potencial aplicación en decisiones críticas.

Tabla 1. Comparación conceptual de técnicas de explicabilidad.

Técnica	Tipo de explicación	Nivel	Ventajas	Limitaciones
SHAP	Basada en teoría de juegos	Local y global	Alta consistencia; interpretaciones robustas	Alto costo computacional en modelos complejos
LIME	Aproximación local con modelos simples	Local	Rápido; flexible; fácil de implementar	Inestabilidad; sensibilidad al muestreo
Modelos sustitutos	Árboles o reglas que imitan al modelo complejo	Global	Fácil de interpretar; útil para auditoría	Pérdida de fidelidad respecto al modelo original
Integrated Gradients	Gradientes acumulados en redes neuronales	Local	Adecuado para modelos profundos	Requiere diferenciabilidad; menos intuitivo para usuarios no técnicos

Los antecedentes investigativos muestran un creciente interés en la aplicación de técnicas de explicabilidad en sectores como la salud, las finanzas, la educación y la administración pública. Estudios recientes evidencian que la explicabilidad contribuye a mejorar la confianza en los sistemas automatizados y facilita la identificación de sesgos que podrían afectar la equidad en la toma de decisiones. Sin embargo, la literatura también señala que existe una brecha en la evaluación comparativa de estas técnicas en contextos institucionales específicos, donde las decisiones críticas requieren altos niveles de transparencia y justificación.

El marco teórico de este estudio se complementa con la consideración de principios éticos y normativos relacionados con la gobernanza algorítmica. Diversas organizaciones internacionales han destacado la importancia de garantizar la transparencia, la responsabilidad y la auditabilidad en el uso de sistemas automatizados. Estos principios orientan la necesidad de adoptar técnicas de explicabilidad que permitan comprender y justificar las decisiones generadas por modelos de aprendizaje automático, especialmente en instituciones donde dichas decisiones tienen implicaciones sociales, administrativas o legales. En conjunto, las teorías, conceptos y antecedentes revisados permiten fundamentar la importancia de evaluar técnicas de explicabilidad en modelos utilizados para decisiones críticas en instituciones. Este marco conceptual orienta el análisis comparativo realizado en el estudio y proporciona las categorías necesarias para interpretar los resultados y valorar su utilidad práctica.

RESULTADOS

El análisis de los modelos de aprendizaje automático entrenados con el conjunto de datos institucional muestra diferencias relevantes en su desempeño predictivo y en la calidad de las explicaciones generadas por las técnicas evaluadas. Los modelos de mayor complejidad, como los algoritmos de gradiente boosting y las redes neuronales, alcanzan los mejores niveles de precisión, aunque presentan mayores desafíos en términos de interpretabilidad. En contraste, modelos más simples como la regresión logística y los árboles de decisión ofrecen menor rendimiento predictivo, pero permiten una comprensión más directa de sus decisiones. Esta variación en la complejidad de los modelos proporciona un escenario adecuado para evaluar la utilidad de las técnicas de explicabilidad seleccionadas.

Las técnicas de explicabilidad aplicadas muestran comportamientos diferenciados en cuanto a fidelidad, estabilidad y tiempo de cómputo. SHAP presenta los valores más altos de fidelidad, reflejando una correspondencia sólida entre las explicaciones generadas y el comportamiento real del modelo. Sin embargo, su tiempo de cómputo es considerablemente mayor, especialmente en modelos complejos. LIME ofrece tiempos de ejecución más rápidos, pero evidencia variabilidad entre ejecuciones, lo que afecta su estabilidad. Los modelos sustitutos muestran un equilibrio entre interpretabilidad y eficiencia, aunque su fidelidad depende de la capacidad del modelo sustituto para aproximar adecuadamente al modelo original. Integrated Gradients presenta un desempeño estable en modelos neuronales, aunque su interpretación resulta menos intuitiva para usuarios no especializados.

Los resultados también evidencian diferencias en la utilidad práctica de las explicaciones para apoyar decisiones críticas. SHAP y los modelos sustitutos generan explicaciones más claras y consistentes para auditores y analistas institucionales, mientras que LIME resulta útil en escenarios donde se requiere rapidez, aunque con menor confiabilidad. Integrated Gradients aporta información relevante en modelos profundos, pero su comprensión requiere mayor formación técnica. En conjunto, los hallazgos permiten identificar fortalezas y limitaciones de cada técnica en función del tipo de modelo y del contexto institucional.

A continuación, se presenta la tabla con los resultados comparativos, en formato compatible con Word.

Tabla 2. Resultados comparativos de las técnicas de explicabilidad

Técnica	Fidelidad	Estabilidad	Tiempo de cómputo	Utilidad para decisiones críticas
SHAP	Alta	Alta	Alto	Muy alta
LIME	Media	Baja	Bajo	Media
Modelos sustitutos	Media	Alta	Bajo	Alta
Integrated Gradients	Alta	Media	Medio	Alta

DISCUSIÓN

Los resultados obtenidos permiten analizar de manera crítica el comportamiento de las técnicas de explicabilidad evaluadas y su contribución a la transparencia en modelos de aprendizaje automático utilizados en decisiones institucionales. La comparación entre técnicas evidencia que no existe un método universalmente superior, sino que su utilidad depende del tipo de modelo, del contexto de aplicación y de las necesidades de interpretación de los usuarios institucionales. Este hallazgo coincide con la literatura especializada, que señala que la explicabilidad debe entenderse como un conjunto de herramientas complementarias más que como una solución única.

La alta fidelidad y estabilidad observadas en SHAP refuerzan su posición como una de las técnicas más robustas para auditoría algorítmica. Su capacidad para asignar contribuciones consistentes a cada variable permite comprender con precisión cómo el modelo genera sus predicciones. Sin embargo, el elevado costo computacional limita su aplicabilidad en escenarios donde se requiere inmediatez o donde los recursos tecnológicos son restringidos. Este contraste entre precisión y eficiencia plantea un desafío para instituciones que deben equilibrar transparencia y operatividad.

LIME, por su parte, muestra ventajas en términos de rapidez y flexibilidad, lo que lo convierte en una herramienta útil para análisis exploratorios o para contextos donde se requiere una explicación inmediata. No obstante, su inestabilidad entre ejecuciones reduce su confiabilidad para auditorías formales o decisiones de alto impacto. Este comportamiento confirma lo señalado por estudios previos, que advierten sobre la sensibilidad de LIME al muestreo y a la variabilidad local del modelo.

Los modelos sustitutos ofrecen un punto intermedio entre interpretabilidad y eficiencia. Su utilidad depende de la capacidad del modelo sustituto para aproximar adecuadamente al modelo original, lo que implica que su fidelidad puede variar según la complejidad del algoritmo principal. Aun así, su claridad conceptual los convierte en una opción valiosa para usuarios no técnicos, especialmente en instituciones donde la comprensión global del modelo es tan importante como la explicación de casos individuales.

Integrated Gradients muestra un desempeño sólido en modelos neuronales, lo que confirma su pertinencia para arquitecturas profundas. Sin embargo, su interpretación requiere conocimientos técnicos avanzados, lo que puede limitar su adopción en instituciones donde los usuarios no poseen formación especializada en aprendizaje profundo.

Este hallazgo subraya la importancia de considerar no solo la calidad técnica de las explicaciones, sino también su accesibilidad para los actores involucrados en la toma de decisiones.

En conjunto, los resultados sugieren que la selección de técnicas de explicabilidad debe responder a criterios situados: tipo de modelo, recursos disponibles, nivel de formación de los usuarios y naturaleza de las decisiones institucionales. La combinación de técnicas emerge como una estrategia recomendable para obtener explicaciones más completas y confiables. Asimismo, los hallazgos refuerzan la necesidad de fortalecer la gobernanza algorítmica mediante prácticas que integren explicabilidad, auditoría y evaluación continua de modelos.

CONCLUSIONES

El estudio permite concluir que la explicabilidad constituye un componente esencial para fortalecer la transparencia, la confianza y la auditoría en el uso de modelos de aprendizaje automático dentro de instituciones. La evaluación comparativa de las técnicas analizadas demuestra que cada una aporta ventajas específicas, pero también presenta limitaciones que deben considerarse al momento de seleccionar la herramienta más adecuada para un contexto determinado. No existe una técnica universalmente superior; más bien, la utilidad de cada método depende del tipo de modelo, del nivel de complejidad algorítmica y de las necesidades de interpretación de los usuarios institucionales.

Los resultados evidencian que SHAP ofrece las explicaciones más consistentes y fieles, lo que lo convierte en una opción sólida para auditorías formales y decisiones críticas. Sin embargo, su alto costo computacional puede limitar su implementación en entornos con recursos tecnológicos restringidos. LIME destaca por su rapidez y flexibilidad, pero su inestabilidad reduce su confiabilidad en escenarios donde la precisión interpretativa es indispensable. Los modelos sustitutos representan una alternativa equilibrada, especialmente útil para usuarios no técnicos, aunque su fidelidad depende de la capacidad del modelo explicativo para aproximar al modelo original. Integrated Gradients muestra un desempeño adecuado en modelos neuronales, pero su interpretación requiere conocimientos especializados, lo que puede limitar su adopción institucional.

En conjunto, los hallazgos subrayan la importancia de adoptar un enfoque estratégico en la selección de técnicas de explicabilidad, considerando factores como el tipo de modelo, la disponibilidad de recursos, el nivel de formación de los usuarios y la naturaleza de las decisiones institucionales.

La combinación de técnicas emerge como una estrategia recomendable para obtener explicaciones más completas y robustas. Asimismo, el estudio destaca la necesidad de fortalecer la gobernanza algorítmica mediante prácticas que integren explicabilidad, evaluación continua y mecanismos de auditoría que garanticen decisiones éticas, transparentes y verificables.

REFERENCIAS BIBLIOGRÁFICAS

Barredo Arrieta, A., & Del Ser, J. (2022). *Explainable artificial intelligence: A practical guide for interpreting machine learning models and deep learning*. Springer.

Burkart, N., & Huber, M. F. (2021). *A survey on the explainability of supervised machine learning*. Springer.

Cisneros Estupiñán, M., & Olave Arias, G. (2012). *Cómo se escribe un artículo científico: Guía para la redacción académica*. Universidad de Antioquia.

Gunning, D., & Aha, D. (2021). *Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models*. Springer.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. En *Advances in neural information processing systems* (pp. 4765–4774).

Molnar, C. (2019). *Interpretable machine learning: A guide for making black box models explainable*. Independently published.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. En *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144).

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (Eds.). (2019). *Explainable AI: Interpreting, explaining and visualizing deep learning*. Springer.

Vilone, G., & Longo, L. (2021). *Explainable artificial intelligence: Foundations, methods and applications*. Springer.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. https://doi.org/10.48550/arXiv.1702.08608