Revista Hambatu Science

Ambato – Ecuador / ISSN 3151-815X (en línea) / octubre - diciembre 2026

Volumen 1, Número 4

https://doi.org/10.63862/rhs-v1n4-382-402-2026

Métodos de inteligencia artificial explicable: revisión de técnicas, aplicaciones y limitaciones

Explainable artificial intelligence methods: a review of techniques, applications, and limitations

Francisco Ricardo Baños Solis
Tecnológico Nacional de México - Instituto Tecnológico de Oaxaca

Maricarmen Montserrat Velázquez Hernández
Tecnológico Nacional de México - Instituto Tecnológico de Oaxaca

Adelina Martínez Nieto
Tecnológico Nacional de México - Instituto Tecnológico de Oaxaca

CC BY

Revista Hambatu Science
Periodicidad: octubre - diciembre
Volumen: 1, Número: 4, Año: 2026

DOI : https://doi.org/10.63862/rhs-v1n4-382-402-2026

Métodos de inteligencia artificial explicable: revisión de técnicas, aplicaciones y limitaciones

Francisco Ricardo Baños Solis
Tecnológico Nacional de México
Instituto Tecnológico de Oaxaca
francisco.solis@itoaxaca.edu.mx
https://orcid.org/0000-0002-0935-9779
México

Adelina Martínez Nieto
Tecnológico Nacional de México
Instituto Tecnológico de Oaxaca
adelina.martinez@itoaxaca.edu.mx
https://orcid.org/0009-0002-9101-7870
México

Maricarmen Montserrat Velázquez Hernández
Tecnológico Nacional de México
Instituto Tecnológico de Oaxaca
maricarmen.mv@itoaxaca.edu.mx
https://orcid.org/0009-0009-5519-3165
México

Recibido: 2026-22-09

Aceptado: 2026-31-10

Publicado: 2026-20-11

© The Author(s). This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Resumen

La inteligencia artificial explicable (XAI) agrupa procedimientos orientados a facilitar el examen de predicciones, representaciones y comportamientos de sistemas de aprendizaje automático. Este artículo analiza las principales familias de técnicas, sus ámbitos de aplicación y sus limitaciones, mediante una revisión narrativa estructurada de publicaciones académicas verificables. Se incorporan elementos de búsqueda y organización documental empleados en estudios bibliométricos, pero no estima indicadores de producción científica ni declara exhaustividad. Se clasifican las propuestas en modelos intrínsecamente interpretables, explicadores locales aproximados, atribuciones basadas en contribuciones, métodos de gradientes, explicaciones contrafactuales y enfoques globales. Los resultados muestran que cada familia responde a una pregunta explicativa distinta y exige criterios particulares de evaluación. LIME aproxima decisiones individuales, SHAP distribuye atribuciones entre variables bajo supuestos definidos, los gradientes caracterizan sensibilidades en redes diferenciables y los contrafactuales identifican modificaciones capaces de alterar resultados. La utilidad de estos recursos se encuentra limitada por inestabilidad, dependencia de distribuciones de referencia, problemas de fidelidad, restricciones causales y divergencias entre evaluación técnica y comprensión humana. Se concluye que una explicación no debe equipararse con validación, causalidad, equidad o confiabilidad del modelo.

Palabras clave: Inteligencia artificial explicable; interpretabilidad; LIME; SHAP; explicaciones contrafactuales; revisión de literatura.

Revista Hambatu Science
Periodicidad: octubre - diciembre
Volumen: 1, Número: 4, Año: 2026

Explainable artificial intelligence methods: a review of techniques, applications, and limitations

Abstract

Explainable artificial intelligence (XAI) comprises methods intended to facilitate scrutiny of predictions, representations, and behaviors of machine-learning systems. This paper examines major methodological families, application areas, and limitations through a structured narrative review of verifiable scholarly publications. The approach draws on bibliographic retrieval and documentary organization practices but neither calculates bibliometric indicators nor claims systematic comprehensiveness. Methods are grouped into inherently interpretable models, local surrogate explanations, contribution-based attributions, gradient approaches, counterfactual explanations, and global techniques. Findings indicate that each family addresses a different explanatory question and therefore requires specific assessment criteria. LIME approximates individual decisions, SHAP allocates feature contributions under explicit assumptions, gradient-based methods identify sensitivities in differentiable networks, and counterfactual methods identify changes that can alter an output. Their usefulness is constrained by instability, reference-distribution dependence, limited fidelity, causal feasibility and disagreement between computational indicators and human understanding. Explanations cannot be treated as direct evidence of model validity, fairness, causality, or trustworthiness.

Keywords: Explainable artificial intelligence; interpretability; LIME; SHAP; counterfactual explanations; literature review.

Introducción

La utilización de sistemas predictivos complejos en decisiones organizacionales, científicas y públicas ha intensificado la necesidad de examinar los fundamentos de sus resultados. Un clasificador puede alcanzar un desempeño elevado en una partición de datos y, sin embargo, operar sobre relaciones espurias, variables sustitutas o patrones que cambian cuando se traslada a otro contexto. La exactitud predictiva mide una propiedad concreta del modelo frente a datos y criterios establecidos; no informa por sí misma sobre las razones que determinan una predicción individual, ni garantiza su estabilidad fuera de la distribución de entrenamiento. Frente a esta separación, la inteligencia artificial explicable se ha desarrollado como un conjunto heterogéneo de técnicas para describir asociaciones utilizadas por sistemas de aprendizaje automático y facilitar su examen crítico (Barredo Arrieta et al., 2020; Schwalbe & Finzel, 2024). En términos de clasificación metodológica, Guidotti et al. (2018) diferencian los enfoques que describen el comportamiento global de los que explican predicciones particulares, mientras Barredo Arrieta et al. (2020) proponen considerar el contexto de uso y los objetivos de transparencia.

En la literatura, el término explicabilidad reúne objetivos que no siempre coinciden. Para quienes desarrollan modelos, una explicación puede ayudar a depurar variables o localizar errores; para profesionales de un dominio, puede apoyar la revisión de una decisión; para personas afectadas, puede comunicar los factores que justifican una salida o señalar alternativas. Estas finalidades plantean exigencias diferentes de granularidad, fidelidad y comunicación. Miller (2019) argumenta, desde las ciencias sociales, que explicar no equivale únicamente a proporcionar una lista de factores, pues intervienen las preguntas del destinatario, la comparación implícita con otros resultados y el contexto de interpretación. Esto impide evaluar toda explicación con una métrica única. Doshi-Velez y Kim (2017) sostienen que la interpretabilidad requiere criterios de evaluación explícitos y no puede inferirse únicamente de que un sistema produzca una visualización. Esta distinción permite examinar las explicaciones como resultados que deben ponerse a prueba.

Las técnicas existentes comprenden explicadores que aproximan una predicción en el entorno inmediato de una observación, métodos que asignan contribuciones numéricas a variables, herramientas que representan regiones relevantes en imágenes y estrategias orientadas a

describir cambios que producirían otra respuesta del sistema. También existen modelos cuya propia estructura permite interpretaciones relativamente directas, como árboles pequeños o modelos aditivos con restricciones. La elección entre estos recursos implica decisiones metodológicas respecto de la unidad explicada, las perturbaciones permitidas, la información accesible del modelo y las condiciones bajo las cuales el procedimiento conserva sentido (Ribeiro et al., 2016; Lundberg & Lee, 2017; Rudin, 2019). La clasificación de Guidotti et al. (2018) ayuda a distinguir procedimientos agnósticos respecto de la arquitectura de aquellos que explotan propiedades internas del modelo. Schwalbe y Finzel (2024) muestran que las taxonomías también difieren en el objeto y el destinatario de la explicación.

A partir de estas tensiones se formula la pregunta que orienta el artículo: ¿qué técnicas de inteligencia artificial explicable identifica la literatura académica, cuáles son sus aplicaciones principales y qué limitaciones deben considerarse al interpretar sus resultados? El objetivo general consiste en caracterizar y contrastar estas familias metodológicas mediante una síntesis crítica de documentos científicos. Como objetivos específicos se propone distinguir las preguntas que responde cada método, reconocer aplicaciones documentadas y organizar los límites técnicos y prácticos que condicionan su utilización. La respuesta se desarrolla en el apartado de resultados y se contrasta en la discusión.

Estado del arte

Interpretabilidad, explicabilidad y naturaleza de la explicación

La interpretabilidad suele referirse a la posibilidad de comprender relaciones internas o reglas de un modelo, mientras que la explicabilidad puede abarcar procedimientos posteriores que describen una salida de un sistema complejo. Esta diferenciación sirve como orientación, pero no constituye una frontera aceptada de manera uniforme. Un modelo lineal ofrece coeficientes directamente inspeccionables; ello no significa necesariamente que su comportamiento sea fácil de comprender cuando incorpora cientos de variables correlacionadas, transformaciones y codificaciones difíciles de interpretar. De manera análoga, una visualización intuitiva puede resultar comunicativamente accesible y a la vez representar de forma insuficiente el comportamiento efectivo de un sistema. La relación entre forma matemática, semántica de las variables y destinatarios resulta, por tanto, decisiva (Molnar, 2022; Schwalbe & Finzel, 2024).

Conviene diferenciar explicación de predicción, descripción estadística y demostración causal. Una atribución elevada indica que, bajo un procedimiento y sus supuestos, cierta variable participa en la salida calculada. No prueba que modificar dicha variable en el mundo real produzca el efecto sugerido. Del mismo modo, una característica destacada en una imagen puede funcionar como señal discriminante para una red sin corresponder a una propiedad necesaria del objeto representado. Los modelos aprenden regularidades observables en datos y pueden depender de variables correlacionadas con procesos omitidos. La comprensión causal requiere supuestos y diseños complementarios, por lo que presentar atribuciones predictivas como causas sería una extrapolación metodológica indebida (Miller, 2019; Wachter et al., 2018). Guidotti et al. (2018) describen la extracción de reglas y los modelos sustitutos como alternativas para aproximar conductas complejas. Sus alcances dependen del grado en que esa representación simplificada conserve el funcionamiento relevante del predictor.

La discusión normativa también influye en la definición de explicabilidad. Rudin (2019) sostiene que, especialmente en contextos de alto riesgo, debe examinarse la posibilidad de utilizar modelos directamente interpretables en lugar de justificar automáticamente una caja negra mediante explicaciones posteriores. Esta posición no implica que todos los problemas admitan modelos sencillos con rendimientos equivalentes. Introduce, más bien, una carga de

justificación: la complejidad adicional debe evaluarse frente al costo de perder trazabilidad. En consecuencia, el diseño de un sistema explicable no comienza cuando ya existe una predicción opaca, sino al seleccionar representaciones, restricciones y criterios de evaluación compatibles con el uso previsto.

Dimensiones para clasificar técnicas explicativas

Los criterios de evaluación propuestos por Doshi-Velez y Kim (2017) separan evaluaciones funcionales, basadas en personas usuarias y basadas en tareas reales, distinción útil para evitar que una métrica computacional se confunda con comprensión humana.

Una primera dimensión distingue técnicas locales y globales. Las primeras consideran una instancia concreta, como la clasificación de un documento o la puntuación de riesgo de un registro; las segundas buscan representar reglas generales, efectos promedio o relaciones dominantes en conjuntos de observaciones. Ambas perspectivas son complementarias, aunque no necesariamente consistentes entre sí. Una explicación local correcta en un entorno restringido puede ofrecer poca información sobre casos distantes, mientras que un resumen global puede ocultar interacciones decisivas para observaciones poco frecuentes. Esta tensión exige indicar explícitamente si los resultados describen un caso, un subconjunto o el funcionamiento agregado del modelo (Molnar, 2022; Schwalbe & Finzel, 2024).

Otra distinción separa métodos agnósticos del modelo y métodos dependientes de una arquitectura específica. LIME puede operar, en principio, con acceso a predicciones de clasificadores diferentes, porque construye aproximaciones locales sin requerir su estructura interna. En contraste, los procedimientos basados en gradientes requieren información de diferenciación y se vinculan particularmente con redes neuronales. Entre ambos extremos se encuentran técnicas con versiones especializadas que mejoran su eficiencia al aprovechar ciertas estructuras, como árboles de decisión. La transferibilidad no es sinónimo de equivalencia: cambiar el tipo de datos, la codificación o la distribución utilizada para generar perturbaciones modifica qué explicación puede construirse (Ribeiro et al., 2016; Lundberg & Lee, 2017; Sundararajan et al., 2017).

Una tercera dimensión se refiere al formato de la evidencia explicativa. Algunos métodos producen pesos, otras reglas, ejemplos representativos, mapas de saliencia, curvas de

dependencia o escenarios alternativos. Los pesos ordenan variables, pero no siempre informan sobre decisiones factibles; las reglas suelen ser fáciles de comunicar, aunque pueden simplificar interacciones; los ejemplos pueden apoyar comparaciones, pero generan dudas sobre representatividad. Evaluar estas salidas únicamente por su apariencia intuitiva puede provocar conclusiones erróneas. La elección de formato debería responder a la pregunta del destinatario y a una prueba de fidelidad respecto del modelo estudiado, no solo a la facilidad de representación gráfica (Nauta et al., 2023).

Metodología

Enfoque y estrategia documental

Se realizó una revisión narrativa estructurada de literatura científica, con alcance descriptivo y analítico. Se recuperaron contribuciones metodológicas originales y revisiones de referencia mediante consulta dirigida en portales de editoriales, repositorios académicos y fichas bibliográficas verificables. El procedimiento tomó de los enfoques bibliométricos la formulación de términos, la identificación de metadatos y la organización de documentos por categorías, pero no desarrolló descarga exhaustiva de registros, deduplicación sobre un conjunto exportado, análisis de redes, indicadores de citación ni diagrama PRISMA. Por ello, los resultados representan una síntesis razonada de las fuentes examinadas, no estimaciones acerca de la totalidad de publicaciones mundiales.

La búsqueda temática se articuló con las expresiones en inglés “explainable artificial intelligence”, “interpretable machine learning”, “local explanations”, “SHAP”, “LIME”, “feature attribution”, “counterfactual explanations”, “saliency maps” y “evaluation of explainable AI”. Se utilizaron combinaciones de estos términos para localizar documentos orientados al desarrollo, clasificación o evaluación de métodos. Se priorizaron publicaciones de 2016 a 2024 por su pertinencia para la consolidación de las principales técnicas, sin interpretar ese intervalo como censo temporal del campo. También se admitieron libros académicos y trabajos previos cuando aportaban definiciones o debates necesarios para la interpretación. En visión por computadora, Selvaraju et al. (2017) desarrollaron Grad-CAM para localizar regiones relevantes en predicciones de redes convolucionales mediante

información de gradientes; el mapa resultante no equivale automáticamente a una explicación causal de la decisión.

La unidad de análisis fue cada documento, no cada autor ni cada cita. Se consideraron como criterios de pertinencia la correspondencia explícita con el objeto de estudio, la posibilidad de verificar título, autoría y lugar de publicación, y el contenido suficiente para identificar mecanismos, aplicaciones o límites. Se excluyeron materiales divulgativos sin trazabilidad bibliográfica y textos cuyos datos editoriales no pudieron corroborarse. Las referencias consultadas no constituyen una muestra aleatoria ni permiten inferir proporciones de técnicas utilizadas en el campo. La selección intencional favorece profundidad conceptual, pero reduce la capacidad de replicar de manera exacta una búsqueda documental no exhaustiva.

Extracción y análisis de información

El análisis se organizó mediante una matriz conceptual con seis campos: familia de técnica, escala local o global, requisitos de acceso al modelo, producto explicativo, aplicaciones documentadas y limitaciones reconocidas. En cada caso se distinguieron afirmaciones de los autores de interpretaciones formuladas para esta revisión. La clasificación emplea ejes conceptuales propuestos en estudios de síntesis previos y permite comparar métodos cuyo funcionamiento matemático no es directamente intercambiable. Así, LIME y SHAP se examinan en relación con la atribución de una predicción, mientras que los contrafactuales se analizan por su capacidad de describir cambios de salida y los modelos intrínsecamente interpretables por sus estructuras verificables (Barredo Arrieta et al., 2020; Schwalbe & Finzel, 2024).

La síntesis siguió tres momentos. Primero se identificó la contribución concreta de cada fuente, evitando tratar revisiones generales como pruebas experimentales de eficacia. Después se agruparon coincidencias y desacuerdos en torno a fidelidad, estabilidad, comprensibilidad, viabilidad y dependencia de supuestos. Finalmente se contrastaron los grupos para responder a la pregunta central. Cuando una publicación propone un procedimiento y exhibe ejemplos de aplicación, se registró su alcance metodológico sin atribuirle un desempeño generalizado en cualquier dominio. No se realizaron experimentos computacionales originales ni se calcularon promedios, correlaciones o pruebas de significación sobre publicaciones.

La calidad argumentativa se controló mediante trazabilidad entre hallazgos y referencias identificables, así como diferenciación entre demostraciones teóricas, ejemplos empíricos y recomendaciones de los autores. La revisión reconoce una limitación relevante: sin estrategias de búsqueda ejecutadas y archivadas dentro de bases bibliográficas específicas, no es posible reproducir un número final de registros encontrados, descartados e incluidos. En consecuencia, no se reportan conteos inventados ni se afirma cumplimiento de PRISMA 2020. Esta transparencia permite describir adecuadamente los alcances del producto y establece las condiciones necesarias para transformarlo, en el futuro, en una revisión sistemática formal. La comparación entre métodos debe considerar que una explicación visualmente coherente puede no resultar fiel al comportamiento del modelo. La propuesta de evaluación de Doshi-Velez y Kim (2017) permite especificar qué dimensión se examina en cada prueba.

Resultados

Explicaciones locales: aproximación y atribución

Las fuentes examinadas permiten reconocer a LIME como un enfoque de explicación local mediante modelos sustitutos. Ribeiro et al. (2016) presentan un procedimiento que construye observaciones perturbadas alrededor del caso de interés, obtiene predicciones del sistema original y ajusta un modelo interpretable ponderando la proximidad de las observaciones. El resultado describe qué variables influyen en la aproximación local, no necesariamente cuáles determinan el comportamiento global de la caja negra. Su aportación principal consiste en permitir el análisis de clasificadores distintos mediante un mecanismo relativamente independiente de su estructura. Sin embargo, la interpretación depende de cómo se define la vecindad, de las perturbaciones generadas y de la capacidad del sustituto para representar la superficie de decisión en ese entorno.

SHAP constituye una familia de explicaciones basadas en contribuciones aditivas, articulada en Lundberg y Lee (2017) a partir de ideas de valores de Shapley. Sus salidas asignan a las variables aportaciones relativas a una predicción frente a una referencia. En términos interpretativos, esto permite distinguir factores que incrementan o disminuyen el valor explicado bajo un sistema definido de asignación. La comparación con LIME muestra que ambos pueden producir explicaciones locales, pero sus premisas no son iguales: uno privilegia

una aproximación local ponderada y otro una asignación de contribuciones sujeta a propiedades de consistencia dentro de un marco teórico. La elección de referencia, el tratamiento de dependencias entre variables y la estrategia de cálculo son parte del significado de una atribución SHAP.

No es metodológicamente correcto deducir que un método es universalmente superior por incorporar una formalización axiomática o una visualización más intuitiva. Un conjunto de pesos puede ayudar a identificar posibles atajos predictivos y, al mismo tiempo, oscurecer relaciones complejas que no se reducen a contribuciones aisladas. En presencia de variables redundantes, las interpretaciones cambian según la manera de representar o distribuir la información compartida. La revisión identifica, por tanto, una condición transversal para explicaciones locales: deben acompañarse de la delimitación del caso, de los parámetros del procedimiento y de pruebas de estabilidad y fidelidad que permitan distinguir una señal reproducible de un artefacto de la técnica (Molnar, 2022; Nauta et al., 2023).

Redes neuronales y explicaciones mediante gradientes

Los métodos de atribución basados en gradientes explotan la diferenciabilidad de redes neuronales para estimar cómo varía una salida respecto de componentes de entrada. Sundararajan et al. (2017) formularon Integrated Gradients a partir de criterios axiomáticos de sensibilidad e invariancia de implementación. En lugar de examinar solamente el gradiente instantáneo de una observación, el método integra gradientes a lo largo de una trayectoria entre una referencia y la entrada que se desea explicar. Este planteamiento reduce algunos problemas asociados a saturación local, aunque su interpretación requiere justificar la selección de un punto de referencia y comprender que la atribución se define sobre una trayectoria matemática concreta. Slack et al. (2020) demostraron que los métodos de explicación post hoc LIME y SHAP pueden ser objeto de ataques diseñados para ocultar comportamientos discriminatorios de un clasificador. Este antecedente impide equiparar la producción de atribuciones con una auditoría de seguridad.

En visión por computadora, las atribuciones suelen representarse como mapas de calor o saliencia sobre píxeles y regiones. Tales representaciones pueden destacar señales utilizadas por una red en la clasificación de imágenes, lo que resulta pertinente para análisis exploratorios,

diagnóstico de fallas o formulación de hipótesis sobre patrones aprendidos. Sin embargo, las regiones resaltadas no constituyen una descripción semántica automática del proceso de reconocimiento. Una imagen atractiva o plausible no demuestra que el mecanismo siga los parámetros aprendidos: puede reflejar propiedades genéricas de bordes, contrastes o estructuras del dato. La separación entre legibilidad visual y sensibilidad del método exige comprobaciones adicionales (Adebayo et al., 2018).

Adebayo et al. (2018) propusieron pruebas de sanidad para identificar si los mapas responden a cambios en parámetros del modelo o a modificaciones de la relación con los datos. Sus hallazgos muestran que algunas representaciones aparentemente informativas pueden mostrar escasa sensibilidad a estos factores. La implicación no es que toda explicación visual carezca de valor, sino que la inspección estética no puede reemplazar evaluaciones controladas. Una revisión de resultados aplicada a imágenes médicas, reconocimiento de objetos o inspección industrial debería documentar qué evidencia permite vincular el mapa con la predicción específica y cuáles son las alternativas técnicas que pueden producir representaciones semejantes.

Explicaciones contrafactuales y modelos interpretables

Las explicaciones contrafactuales responden a una pregunta distinta de la asignación de importancia: qué condiciones tendrían que cambiar para obtener otra salida del sistema. Wachter et al. (2018) examinan su pertinencia para decisiones automatizadas y discuten procedimientos que construyen escenarios alternativos sin revelar necesariamente la estructura interna de la caja negra. Si un modelo clasifica una observación de cierta manera, un contrafactual busca otra observación próxima que obtenga una predicción diferente. La cercanía debe entenderse dentro de una métrica específica; dos situaciones pueden estar próximas matemáticamente y, sin embargo, requerir transformaciones incompatibles con las restricciones institucionales o físicas del entorno.

La utilidad de un contrafactual depende de su plausibilidad, diversidad y posibilidad de implementación. Un escenario que sugiera alterar variables inmodificables, como características históricas registradas, sería poco accionable, aunque lograra cambiar la clasificación. Asimismo, la existencia de un cambio matemático en las entradas no garantiza

que la intervención cause el resultado deseado en la realidad. Esto exige distinguir intervención causal, asociación aprendida por el modelo y recomendación normativa. Los métodos contrafactuales son particularmente valiosos para construir preguntas comprensibles por personas no especializadas, pero requieren reglas de factibilidad, restricciones del dominio y revisión humana antes de convertirse en orientaciones operativas (Wachter et al., 2018; Miller, 2019).

Los modelos intrínsecamente interpretables constituyen una alternativa conceptual frente a las explicaciones posteriores. Árboles pequeños, reglas transparentes y determinadas especificaciones aditivas permiten inspeccionar directamente relaciones utilizadas para producir un resultado. Rudin (2019) enfatiza esta posibilidad en contextos de alto riesgo y objeta el supuesto de que necesariamente debe sacrificarse desempeño para ganar interpretabilidad. Su planteamiento debe leerse como un programa de evaluación comparativa, no como una afirmación de que todo árbol sencillo supera a cualquier red profunda. Un diseño responsable contrasta modelos candidatos con criterios de capacidad predictiva, complejidad, estabilidad, legibilidad y costos de error, antes de decidir si se justifica una caja negra.

Aplicaciones y pertinencia por dominio.

El estudio de Selvaraju et al. (2017) también muestra aplicaciones de Grad-CAM en clasificación de imágenes, generación de descripciones y respuesta visual a preguntas, pero su evidencia no autoriza extrapolar indistintamente esas propiedades a cualquier arquitectura.

La literatura identifica aplicaciones de XAI en clasificación textual, reconocimiento de imágenes, análisis clínico, sistemas de decisión y tareas científicas. Ribeiro et al. (2016) ilustran explicaciones de clasificadores de texto e imagen; Sundararajan et al. (2017) trabajan con ejemplos de imágenes, texto y química; Rudin (2019) discute las implicaciones de modelos opacos para decisiones de consecuencias elevadas. Estos trabajos muestran diversidad de uso, pero no autorizan a concluir que una técnica preserve idénticas propiedades en todos los sectores. Los objetivos de validación difieren si la explicación se orienta a depurar un predictor, examinar patrones de datos, comunicar un resultado a una persona afectada o justificar una intervención institucional.

En problemas de datos estructurados, los métodos de contribución por variables ofrecen una interfaz analítica familiar para equipos que trabajan con regresiones y modelos predictivos. Sin embargo, deben interpretarse con particular cautela cuando existen colinealidad, variables derivadas y categorías con codificaciones complejas. En texto, el ocultamiento o sustitución de palabras puede generar frases que no pertenecen a la distribución lingüística utilizada durante el entrenamiento. En imágenes, las máscaras o perturbaciones pueden introducir patrones artificiales. La comparación transversal muestra que la validez de una explicación depende, además del algoritmo, del mecanismo mediante el cual se alteran datos para interrogar al modelo (Molnar, 2022).

Las aplicaciones de alto riesgo demandan una evaluación que excede el rendimiento computacional. Un sistema puede generar explicaciones fieles a sus propias reglas y, al mismo tiempo, utilizar variables relacionadas con sesgos estructurales. Una explicación clara puede facilitar la detección del problema, pero no resolverlo. A su vez, revelar ponderaciones o ejemplos contrafactuales no sustituye la comprobación de calibración, los análisis de errores por grupos, las políticas de protección de datos o la rendición de cuentas. De acuerdo con la lectura conjunta de Rudin (2019), Barredo Arrieta et al. (2020) y Miller (2019), la explicabilidad debe insertarse dentro de un procedimiento de validación y gobernanza, evitando tratarla como una certificación autónoma del sistema.

Limitaciones y evaluación de las explicaciones

La primera limitación transversal identificada es la fidelidad. Una salida explicativa puede describir de forma incompleta cómo opera el modelo original, en particular cuando utiliza una aproximación local o simplifica interacciones. La fidelidad debe examinarse respecto de preguntas precisas: si una regla reproduce predicciones en el vecindario definido, si una atribución cambia cuando cambia el modelo o si un contrafactual realmente cruza la frontera decisoria prevista. La idea de “buena explicación” no puede depender exclusivamente de su comprensión inmediata. Nauta et al. (2023) organizan criterios conceptuales para evaluar XAI y evidencian que la literatura utiliza propiedades heterogéneas; esto fundamenta la necesidad de declarar el criterio aplicado, en lugar de presentar una calificación indiferenciada de calidad.

La segunda limitación es la estabilidad. Cuando explicaciones cambian de manera pronunciada ante pequeñas perturbaciones de entradas, semillas aleatorias o configuraciones, su uso para justificar una decisión puede ser problemático. No toda variación implica un defecto: modelos con fronteras de decisión abruptas pueden producir explicaciones distintas en regiones genuinamente diferentes. La cuestión relevante consiste en determinar si el cambio explicativo responde al sistema que se quiere comprender o a decisiones del procedimiento de explicación. Para abordar este problema se requieren repeticiones, análisis de sensibilidad, reportes de intervalos o medidas de dispersión, sin confundir estabilidad del algoritmo de explicación con estabilidad del predictor.

La tercera limitación remite a la relación entre métricas técnicas y comprensión humana. Miller (2019) señala que las personas suelen solicitar respuestas selectivas, contrastivas y contextualizadas. Una explicación que reproduzca con fidelidad matemática cierto resultado puede carecer de utilidad si utiliza variables ininteligibles o presenta una cantidad excesiva de factores. La evaluación con personas usuarias no debe limitarse a preguntar si una gráfica les parece convincente: debe examinar qué decisiones permite tomar, qué errores induce y si ayuda a identificar situaciones en las que no conviene confiar en el sistema. Nauta et al. (2023) refuerzan la necesidad de evaluación multifacética, evitando usar popularidad visual como equivalente de evidencia.

Síntesis de resultados y respuesta a la pregunta de investigación

La pregunta de investigación puede responderse en tres componentes. Primero, las técnicas reconocidas en la literatura incluyen modelos intrínsecamente interpretables, aproximaciones locales como LIME, atribuciones aditivas como SHAP, gradientes integrados y mapas de saliencia en redes neuronales, así como explicaciones contrafactuales. No constituyen variantes de una única operación: algunos recursos describen asociaciones de variables con predicciones, otros aproximan el funcionamiento cerca de una observación y otros identifican escenarios alternativos compatibles con una salida distinta. La elección de técnica debe partir de la pregunta que se requiere responder y del tipo de modelo disponible (Ribeiro et al., 2016; Lundberg & Lee, 2017; Sundararajan et al., 2017; Wachter et al., 2018). La revisión de Guidotti et al. (2018) identifica métodos de reglas, ejemplos, atribuciones y aproximaciones sustitutas;

confrontarla con Schwalbe y Finzel (2024) permite reconocer que las taxonomías responden a criterios diferentes y no establecen una jerarquía universal.

Segundo, las aplicaciones más claramente documentadas en las fuentes revisadas corresponden a la explicación de clasificadores de texto e imagen, la interpretación de modelos predictivos tabulares, la exploración de redes neuronales y el análisis de decisiones automatizadas. La transferibilidad entre dominios requiere condiciones adicionales. Una atribución calculada para variables tabulares no ofrece de manera automática una explicación útil en lenguaje natural; una saliencia visual no constituye prueba clínica; y una ruta contrafactual no garantiza que el cambio sugerido sea factible. Las aplicaciones permiten ejemplificar capacidades metodológicas, pero no proporcionan, por sí solas, evidencias para establecer jerarquías universales entre técnicas.

Tercero, los principales límites se agrupan en fidelidad insuficiente, sensibilidad a referencias y perturbaciones, inestabilidad, problemas de causalidad y factibilidad, diferencias entre validez matemática y comprensión humana, y riesgos de interpretar la explicación como garantía de justicia o seguridad. La respuesta sintética es que ninguna familia explica por sí sola todas las propiedades relevantes de un sistema predictivo. La combinación de modelos transparentes, análisis locales y globales, pruebas de sensibilidad y evaluación del contexto puede mejorar la calidad de la inspección, pero solo si sus productos se distinguen explícitamente de las propiedades que no verifican (Rudin, 2019; Nauta et al., 2023; Schwalbe & Finzel, 2024).

Discusión

Los hallazgos permiten reconocer un desplazamiento del debate desde la disponibilidad de herramientas hacia la evaluación de sus resultados. Los primeros métodos de explicación local ofrecieron procedimientos operativos para examinar clasificadores complejos y ampliaron las posibilidades de inspección técnica (Ribeiro et al., 2016; Lundberg & Lee, 2017). Los trabajos posteriores que problematizan las medidas de calidad muestran que producir una explicación no equivale a demostrar que sea correcta o útil (Adebayo et al., 2018; Nauta et al., 2023). Estas contribuciones no son contradictorias: responden a etapas distintas del problema computacional. La primera construye formas de observación y la segunda examina cuándo esas formas proporcionan evidencia informativa. Además, el resultado de Slack et al. (2020)

problematiza una premisa habitual: la disponibilidad de explicaciones no garantiza que estas revelen prácticas discriminatorias, si el propio mecanismo de explicación puede ser manipulado.

También se observa una tensión entre la búsqueda de generalidad y las condiciones específicas de cada dominio. Los métodos agnósticos facilitan su implementación sobre sistemas heterogéneos, aunque deben crear representaciones auxiliares para interrogar al modelo. Estas representaciones pueden afectar la fidelidad de la explicación. Por su parte, los métodos que aprovechan estructuras particulares pueden describir propiedades computacionales mejor delimitadas, pero ofrecen menor portabilidad. Schwalbe y Finzel (2024) señalan la diversidad de criterios de clasificación existentes; esto ayuda a entender por qué comparar técnicas únicamente por denominación o frecuencia de uso resulta insuficiente. La comparación debe considerar objetivos explicativos, supuestos operativos y formas de validación.

La postura de Rudin (2019) introduce una crítica de diseño: una explicación poshoc no elimina los problemas de gobernanza y evaluación de un sistema opaco. Este argumento cobra particular relevancia cuando las decisiones tienen consecuencias difíciles de revertir. Sin embargo, sería excesivo concluir que la interpretabilidad estructural es una solución automática. Los modelos transparentes también pueden entrenarse con datos defectuosos, producir errores desiguales o depender de variables mal definidas. La contribución de la crítica consiste en cuestionar la elección automática de complejidad y en exigir comparaciones explícitas entre candidatos, no en reemplazar la evaluación empírica por una preferencia universal hacia la sencillez. La distinción entre evaluación de laboratorio y adecuación a una tarea concreta, formulada por Doshi-Velez y Kim (2017), refuerza la necesidad de documentar perfiles de destinatarios, condiciones de uso y límites de la evidencia.

Desde la interacción humano-computadora, el análisis requiere desplazar el centro de atención de los atributos del método hacia la relación entre explicación y destinatario. Miller (2019) ayuda a advertir que la suficiencia explicativa no tiene un significado único: una persona encargada de depurar código, una profesional de salud y una persona que recibe una decisión automatizada pueden formular preguntas diferentes sobre el mismo resultado. Para producir explicaciones responsables, es necesario documentar la tarea de uso, el nivel de conocimiento,

las posibles interpretaciones erróneas y la evidencia que justifica cada afirmación. Esta perspectiva complementa la evaluación de fidelidad y robustez, pero no la sustituye.

Conclusiones

La revisión permite concluir que la inteligencia artificial explicable constituye un campo de métodos diversos y no un procedimiento único para descubrir el funcionamiento de una caja negra. LIME, SHAP, gradientes integrados y contrafactuales responden a preguntas distintas: aproximación local, distribución de contribuciones, sensibilidad diferencial y modificación hipotética de resultados, respectivamente. Los modelos intrínsecamente interpretables agregan la opción de inspeccionar directamente la estructura decisoria. Así, el objetivo de explicación debe definirse antes de seleccionar una herramienta y debe estar vinculado con las características de los datos, del algoritmo predictivo y del contexto de aplicación.

La aportación analítica del artículo es una propuesta de lectura conjunta en tres planos: alcance de la explicación, requisitos técnicos y calidad de la evidencia. El primer plano identifica si el procedimiento describe un caso, una relación global o un escenario alternativo. El segundo reconoce la dependencia de parámetros, referencias, perturbaciones y acceso a la estructura interna. El tercero exige comprobar fidelidad, estabilidad, coherencia con el contexto y utilidad para quienes emplearán la salida. Esta lectura impide trasladar resultados de una dimensión a otra sin comprobación y ofrece criterios para diseñar evaluaciones comparativas posteriores.

El estudio tiene limitaciones específicas. La selección de fuentes fue intencional y no exhaustiva, no existe un conjunto de registros bibliográficos exportados y depurados, y no se aplicaron procedimientos de metaanálisis ni experimentos originales. Por tanto, no es posible afirmar cuáles técnicas dominan cuantitativamente las publicaciones ni medir su rendimiento comparativo. Como continuidad se propone ejecutar búsquedas reproducibles en Scopus y Web of Science, registrar criterios de elegibilidad, evaluar la calidad de los estudios empíricos y contrastar métodos mediante conjuntos de datos públicos y protocolos homogéneos. Esas etapas permitirían pasar de una síntesis conceptual a pruebas de afirmaciones específicas sin confundir descripción documental con inferencia causal o validación computacional.

Referencias bibliográficas

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. En Advances in Neural Information Processing Systems (Vol. 31). https://arxiv.org/abs/1810.03292

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

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

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), Article 93. https://doi.org/10.1145/3236009

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. En Advances in Neural Information Processing Systems (Vol. 30). https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable (2.ª ed.). https://christophm.github.io/interpretable-ml-book/

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., & Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. ACM Computing Surveys, 55(13s), artículo 295. https://doi.org/10.1145/3583558

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). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1, 206–215. https://doi.org/10.1038/s42256-019-0048-x

Schwalbe, G., & Finzel, B. (2024). A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. Data Mining and Knowledge Discovery, 38, 3043–3101. https://doi.org/10.1007/s10618-022-00867-8

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. En Proceedings of the IEEE International Conference on Computer Vision (pp. 618–626). https://doi.org/10.1109/ICCV.2017.74

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. En Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (pp. 180–186). https://doi.org/10.1145/3375627.3375830

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. En Proceedings of the 34th International Conference on Machine Learning (Vol. 70, pp. 3319–3328). Proceedings of Machine Learning Research. https://proceedings.mlr.press/v70/sundararajan17a.html

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841–887. https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf

Declaraciones finales

Conflicto de intereses: El autor declara no tener ningún conflicto de interés relacionado con el presente trabajo.

Financiamiento: Esta investigación no recibió financiamiento externo de agencias públicas, comerciales o sin fines de lucro.

Agradecimiento: Los autores agradecen al Tecnológico Nacional de México Campus Oaxaca por el respaldo académico e institucional brindado para el desarrollo de esta investigación. Al primer autor por la dirección del trabajo de investigación y al resto de autores por sus aportes y la investigación de campo.

Uso de inteligencia artificial generativa: NA.

Nota editorial: El presente artículo es original e inédito y no es producto de una publicación anterior. Dirección de la investigación - Francisco Ricardo Baños Solis. Elaboración de la investigación: Emilio Raúl Peña Pérez, Mario Sergio Lerín Cruz, Maria Fernanda Cortes Martinez, Manuel de Jesús Melo Monterrey.