e-ISSN 2594-1682 Sintaxis, año 9, núm. 17, julio-diciembre, 2026

Más allá de la opacidad: marco conceptual de los modelos interpretables y enfoques de inteligencia artificial explicable (XAI)

Beyond Opacity: A Conceptual Framework for Interpretable Models and Approaches to Explainable Artificial Intelligence (XAI)

Juan David Luján-Villar*
Secretaría de Educación del Distrito
Av. El Dorado # 66–6, El Salitre-Suba (IED),
Sede A, Bogotá, Colombia

lujanvillar@gmail.com
https://orcid.org/0000-0001-8622-4774

Editor invitado: Dr. Ignacio Quepons Ramírez

https://doi.org/10.36105/stx.2026n17.02

Fecha de Recepción: 20 de abril de 2026
Fecha de Aceptación: 19 de mayo de 2026

RESUMEN

Este trabajo propone la realización de una comparación entre modelos interpretables (*interpretable models*) y las técnicas de explicación (*explanation techniques* o *XAI methods*) en términos taxonómicos de demarcación conceptual. A través de un enfoque denominado desarrollo de marco conceptual (DMC) y la aplicación formal de un modelo de este campo, se ejemplifica el uso de esta tecnología disponible respecto al campo de las ciencias sociales y humanas. Los resultados evidencian un enorme potencial de esta área en la aplicación metodológica de estos enfoques en las ciencias sociales y humanas de vanguardia. Se concluye con la reflexión pertinente del aporte computacional de estas ideas sobre problemas complejos en términos

* Doctorando en Docencia y Educación Artística, Universidad de las Américas y el Caribe, UNAC (Colima, México), Magister en Investigación Social Interdisciplinaria y Licenciado en Educación artística (UDFJC) (Bogotá, Colombia) y docente de la Secretaría de Educación del Distrito (Bogotá, Colombia).

CÓMO CITAR: Luján-Villar, J. D. (2026). Más allá de la opacidad: marco conceptual de los modelos interpretables y enfoques de inteligencia artificial explicable (XAI). *Sintaxis, año 9, núm. 17*, DOI: https://doi.org/10.36105/stx.2026n17.02

de modelado, datos, evidencias y la transparencia necesaria de las dinámicas investigativas contemporáneas.

Palabras clave: inteligencia artificial explicable, modelos interpretables, aprendizaje automático, transparencia.

ABSTRACT

This paper proposes a comparison between interpretable models and explanation techniques (or XAI methods) in terms of taxonomic conceptual demarcation. Through an approach known as conceptual framework development (CFD) and the formal application of a model from this field, the paper illustrates the use of this available technology in the context of the social sciences and humanities. The results demonstrate the enormous potential of this area in the methodological application of these approaches in cutting-edge social sciences and humanities. The paper concludes with a pertinent reflection on the computational contribution of these ideas to complex problems in terms of modeling, data, evidence, and the necessary transparency of contemporary research dynamics.

Keywords: explainable artificial intelligence, interpretable models, machine learning, transparency.

INTRODUCCIÓN

Existe un debate contemporáneo notable entre los modelos interpretables (*interpretable models*) y las técnicas de explicación (*explanation techniques* o *XAI methods*) dónde se percibe una posible confusión fundamental en la investigación actual. Ambos conceptos abordan el problema de la opacidad algorítmica y la necesidad de *transparencia* en las lógicas internas del modelado, debido a que operan bajo supuestos teóricos distintos, producen representaciones epistemológicas diferentes y en algunos casos son complementarias. Los modelos interpretables son aquellos cuya lógica interna es comprensible por sí misma o por diseño, como los árboles de decisión (*decision trees*). Por otra parte, las técnicas de explicación son herramientas externas que se aplican sobre modelos complejos para hacer comprensibles los resultados *a posteriori*.

Una opción que emana de esta disyuntiva es usar modelos intrínsecamente interpretables y aplicar técnicas de explicabilidad a modelos complejos, para lograr transparencia en sistemas de aprendizaje automático (Hsieh *et al.*, 2024). El presente estudio tiene como

objetivo realizar de una taxonomía de los modelos interpretables y las técnicas XAI de manera simultánea a la aplicación de ejercicios formales que incluyen ejecución algorítmica y modelado, con el ánimo explicativo de evidenciar las posibilidades que brindan los usos de estas tecnologías en las diversas ciencias humanas, sociales y formales más allá de algún lineamiento teórico específico (De *et al.*, 2024; Dib & Capus, 2025).

METODOLOGÍA

El método que se emplea es el desarrollo del marco conceptual (DMC) (Jabareen, 2009) basado en una presentación taxonómica fundamentada en cinco controles metodológicos integrados: taxonomía, delimitación, caracterización, verificabilidad documental y demarcación epistémica. A propósito de los modelos interpretables y las técnicas XAI (Hsieh *et al.*, 2024; Kalasampath *et al.*, 2025), en consonancia con la aplicación de técnicas algorítmicas se define la aplicabilidad y alcances de estas prestaciones. Para ejemplificar el uso y posibilidades tecnológicas de estos enfoques se crea y ejecuta un modelo XAI basado en dos técnicas canónicas: SHAP (*shapley additive explanations*) y LIME (*local interpretable model-agnostic explanation*) en un caso concreto de diagnóstico clínico de salud social, ciencia de datos y aprendizaje automático (*machine learning*, ML). No se abordan los aspectos genealógicos de estos campos debido a que escapa de las posibilidades de este espacio. El carácter de este enfoque es propositivo, delimita y caracteriza el objeto de estudio: los modelos interpretables y las principales técnicas XAI respecto a la implementación en ciencias sociales.

La aplicación del método DMC se realiza mediante la propuesta de Jabareen (2009) con la resolución del mapeo de fuentes, lectura y categorización que precisa la elección documental según criterios de selección bibliográfica como relevancia temática, la actualidad de los enfoques del campo de investigación analizado y la identificación de conceptos. Durante la consulta de información se acude a las principales fuentes de registros científicos como Scopus, Web of Science y OpenAlex. Además, se realizan indagaciones específicas en las plataformas Google Académico, Crossref y arXiv. Después se examinan los argumentos y metodologías de manera detallada con la finalidad de comprender cómo están construidas las ideas centrales, más allá de la simple aceptación del significado. Los conceptos se integran y relacionan de manera lógica y coherente con el debido soporte académico para realizar una síntesis integral en tablas e imágenes, validación soportada en revisión y contraste. Finalmente, se efectúa un proceso de reinterpretación en la taxonomía propuesta y las aplicaciones de modelado empírico correspondientes donde se identifica un problema, se indaga la literatura correspondiente, se evalúan y analizan los datos y se presentan brevemente de los resultados según el programa de revisión integradora y teórica propuesto por Whittemore & Knafl (2005).

Para la aplicación del modelado empírico se crea un código en el lenguaje de programación Python 3.13.2, estructurado en seis fases a través de ML con técnicas XAI aplicado a una base datos de carácter étnico. Esta aplicación tiene como objetivo combinar las técnicas SHAP y LIME para interpretar las predicciones de un modelo práctico de clasificación de diabetes. 1. Se configuran las librerías y la reproducibilidad al fijar una semilla aleatoria (SEED=42) en todos los generadores NumPy y el sistema operativo. 2. Se realiza el preprocesamiento de datos, donde se reemplazan con NaN los ceros fisiológicamente imposibles en columnas como glucosa o presión arterial, y se imputan dichos valores con la técnica KNNImputer (5 vecinos), y se dividen los datos en conjuntos de entrenamiento y prueba (80/20) con escalado StandardScaler. 3. Se optimiza un RandomForestClassifier mediante GridSearchCV para la validación cruzada estratificada de 5 pliegues, con el fin de evaluar el modelo resultante con métricas de *accuracy*, F1, AUC-ROC, *recall* y *precision* e identificar el umbral de decisión óptimo. 4. Se seleccionan 5 instancias representativas del conjunto de prueba: alta confianza en diabetes, alta confianza en no-diabetes, e incertidumbre para explicaciones locales. 5. Se generan explicaciones LIME que tienen hasta tres rondas de muestreo progresivo (10k, 20k y 50k muestras) para mejorar la fidelidad local de cada explicación, y explicaciones SHAP lo que produce gráficos de cascada, dependencia y fuerza. Finalmente, se calcula la concordancia entre SHAP y LIME por presencia y posición en el top-4 de variables (*features*) de la base de datos.

INTELIGENCIA ARTIFICIAL EXPLICABLE (XAI)

Pretolesi *et al.* (2025) consideran que la IA posibilita incluir en proyectos de investigación definiciones, encuadres, modelado y apoyo en los aspectos centrales de la resolución de problemas. De manera general, las técnicas XAI buscan establecer transparencia, interpretabilidad y explicabilidad en los modelos inteligencia artificial (IA) para la investigación, con múltiples enfoques que implican explicaciones *post-hoc* es decir, descripciones detalladas después del entrenamiento de modelos, métodos de transparencia en la modelación y técnicas de visualización interactivas o tradicionales, con la finalidad de esclarecer el comportamiento de los modelos y las opciones que brindan los sistemas de IA (Mathew *et al.*, 2025).

Kalasampath *et al.* (2025) reconocen las técnicas SHAP y LIME como las más relevantes en lo que se denomina aplicaciones XAI, al interior de campos como la salud, el diagnóstico y la imagenología médica, automóviles, robótica, gestión ambiental y agrícola, optimización en el campo industrial, ciberseguridad sistemática, finanzas e inversiones, el sistema legal y regulatorio, comercio electrónico (*eCommerce*), transporte y redes sociales de segunda generación

(Aluvalu *et al.*, 2024; Gaur & Abraham, 2024; Juarez *et al.*, 2024; Khamparia & Gupta, 2025; Kumar *et al.*, 2025; Malviya & Sundram, 2025; Saarela & Podgorelec, 2024). Los autores reconocen en las técnicas y modelos XAI principios de una estabilidad sólida y ciertas garantías matemáticas, debido a la transparencia de estas arquitecturas en términos de modelado algorítmico.

Continuamente la IA es criticada por la recurrencia de modelos de caja negra (*black-box models*) (Grobrügge *et al.*, 2024), aplicaciones que incluyen entradas y salidas, lógicas internas, estadísticas, recursividades y procesamiento de datos aspectos que son evidentemente opacos por diseño. Algunas técnicas y modelos XAI intentan proporcionar modelos de IA de caja blanca (*white-box models*) como respuesta al modelado opaco. De este razonamiento proviene el término *explicable* o IA de caja de cristal como algo opuesto a los modelos convencionales de la IA no transparentes e inherentemente oscuros (Kalasampath *et al.*, 2025). Diversos modelos son señalados por falta de nitidez: redes neuronales profundas (*deep learning*), bosques aleatorios (*random forests*), modelos de *boosting* como XGBoost o máquinas de soporte vectorial (*SVM*) (Joyce *et al.*, 2023), la interpretabilidad de estas metodologías puede tornarse técnica y tortuosa en extremo. Algo similar ocurre con los grandes modelos de lenguaje (*large language models*, LLMs) de la actualidad (Zhang *et al.*, 2025).

Hsieh *et al.* (2024) presentan un inventario completo y actual de las técnicas XAI y los algoritmos de aplicación. Los criterios que brindan los modelos y técnicas obedecen a cuatro principios fundamentales; 1) deben ser *equitativos* y no brindar resultados sesgados de tipo discriminativo, sobre atributos sensibles de la condición humana como geolocalización, aspectos físicos, edad o género; 2) *transparentes* en el acceso de la estructura y el manejo de los datos del modelo; 3) *comprensibles* en los mecanismos internos de los modelos ML; y 4) *interpretables* en la comprensión humana de una acción decisiva. Es posible clasificar las técnicas y aplicaciones XAI de maneras diversas según el reciente índice general ofrecido por Hsieh *et al.* (2024) que aquí se amplía y complementa:

1. Modelos tradicionales de aprendizaje automático (ML): son algoritmos estadísticos no neuronales que aprenden mapeo de entrada y salida a partir de datos, optimización paramétrica o inductiva sobre una función matemática explícita que puede ser lineal, aditiva o basada en árboles. Operan bajo la premisa de generalización estadística a partir de patrones discernibles y pueden no asumir estructuras jerárquicas internas complejas en algunos casos (Chakraborty *et al.*, 2021; Huang *et al.*, 2024; Liu *et al.*, 2022; Murdoch *et al.*, 2019; Rudin *et al.*, 2022). Las técnicas clásicas en este campo son: árboles de decisión, regresiones lineales (*linear regression*), máquinas de vectores soporte (*support vector machines,* SVM), modelos

bayesianos (*bayesian models*), k-vecinos más cercanos (*k-nearest neighbors*), aprendizaje automático basado en reglas (*rule-based machine learning*¸ GAMs), modelos generalizados aditivos (*generalized additive model*), y redes neuronales multicapas (*multi–layer neural networks*) (Barredo *et al.*, 2020).

2. Modelos de aprendizaje profundo: son arquitecturas computacionales neuronales profundas que operan en un espacio de alta dimensión, y utilizan funciones de activación no lineales y múltiples capas para aprender representaciones jerárquicas automáticas de los datos. Pueden tener capacidad de invariancia a transformaciones y dependencia del gran volumen de datos (*big data*) como tendencia general. En este grupo se localizan redes neuronales convolucionales (*convolutional neural networks*, CNNs), modelos basados en redes neuronales recurrentes (*recurrent neural networks*, RNN), técnicas como la visualización de características (*feature visualization*) (Jena, 2023; Rehman & Saba, 2025) y mecanismos de atención (*attention mechanisms*) (Grobrügge *et al.*, 2024).
3. Los grandes y reconocidos modelos lingüísticos (LLMs): son modelos estadísticos masivos pre-entrenados que utilizan arquitecturas *transformer* y técnicas de predicción autorregresiva (*autoregressive prediction*) o predicción del siguiente token (*next-token prediction*), con billones o trillones de parámetros, entrenados sobre corpus multilingües masivos para generalizar tareas de procesamiento de lenguaje natural derivadas de inferencia probabilística y simulación de razonamiento. Algunos modelos conocidos son: BERT, GPT y T5, entre los cuales son notorias las técnicas de sondeo (*for probing*), análisis basado en gradientes (*gradient-based analysis*), e interpretación del peso de la atención (*attention weight interpretation*). Existen también modelos LLMs entrenados de manera específica sobre aspectos científicos puntuales, estos modelos se clasifican como grandes modelos lingüísticos científicos (*scientific large language models*, *sci-LLMs*), grandes modelos lingüísticos científicos textuales (*textual scientific large language models*, text-sci-LLMs) o grandes modelos multimodales de lenguaje científico (*multimodal scientific large language models, MM-Sci-LLMs*) (Zhang *et al.*, 2025).
4. Entre las técnicas destacables de la IA explicable se encuentran: técnicas para la interpretación de modelos (*model interpretation*), métodos intrínsecos (*intrinsic methods*) establecidos en la importancia de las características, métodos *post-hoc* entre los cuales se encuentran los reconocidos SHAP, LIME (Lundberg & Lee, 2018) y Grad-CAM, explicaciones contrafactuales (*counterfactual explanations*) y las técnicas de inferencia causal (*causal inference techniques*) (Molnar, 2022). Son métodos algorítmicos diseñados para aproximar, descomponer o revelar la lógica interna de

modelos opacos, al proporcionar representaciones interpretables del comportamiento predictivo.

5. Aplicaciones interdisciplinares: son análisis de casos prácticos desde diversos campos y sectores como salud, sanidad, finanzas e inversiones, derecho y formulación de políticas en diversos sectores (Kumar *et al.*, 2025; Malviya & Sundram, 2025; Saarela & Podgorelec, 2024). Estos desarrollos demuestran formas posibles de toma de decisiones y abordan diversas problemáticas de equidad. Se pueden definir como una transposición de capacidades computacionales a dominios fuera de la informática, para resolver problemas complejos que requieren inferencia, predicción o análisis de patrones en entornos de datos estructurados o no estructurados.
6. Evaluación y retos: se realizan a partir de debates detallados sobre la evaluación de la calidad de las explicaciones mediante métricas como fidelidad (*fidelity*), estabilidad (*stability*) y comprensibilidad (*comprehensibility*). Cuestiona dificultades de la naturaleza de los modelos de caja negra, modelos profundos (*deep models*) e intersecciones entre precisión (*accuracy*), flexibilidad (*flexibility*) e interpretabilidad (*interpretability*). Este proceso crítico de medir la calidad de los sistemas no solo desde la precisión predictiva (exactitud o F1), además aborda robustez, equidad, estabilidad e interpretabilidad. Los principales retos incluyen la compensación (*trade-off*) entre rendimiento y explicabilidad, sesgos inherentes a los datos de entrenamiento, y la dificultad de garantizar que una explicación sea causal y no meramente correlacional.
7. Herramientas y marcos: son conjuntos de bibliotecas, repositorios, entornos de desarrollo y estándares abiertos que facilitan el despliegue, entrenamiento, visualización y análisis de modelos para abordar marcos de trabajo XAI, herramientas agnósticas de modelos (*model-agnostic tools*) como SHAP y LIME (Dafali *et al.*, 2023; Lundberg & Lee, 2018), bibliotecas de aprendizaje profundo como Captum (Grobrügge *et al.*, 2024), y marcos de visualización (*visualization frameworks*) de explicaciones interactivas generalmente a través de plataformas especializadas.

APLICACIONES

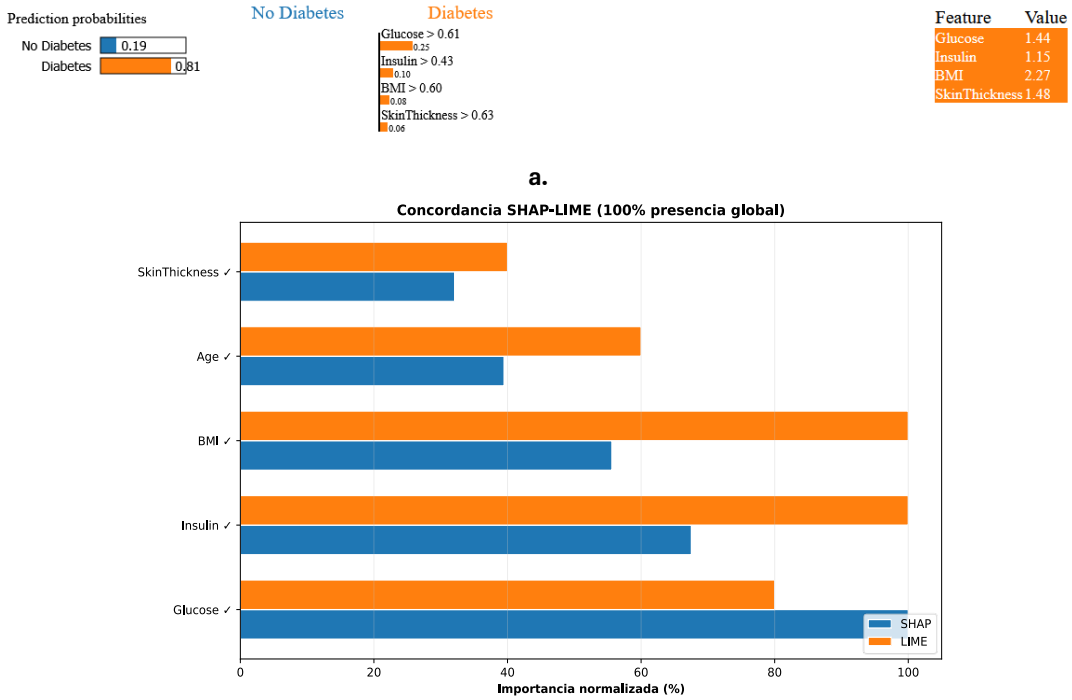
La Figura 1 presenta la distribución de predicciones del modelo RandomForest optimizado con los parámetros: *max depth*=4, *n estimators*=200, *umbral*=0.35, sobre el conjunto de prueba

denominado Pima Indians Diabetes[1] (*n*=154), una base de datos real de naturaleza étnica sobre posibles casos de pacientes con diabetes. El modelo clasifica al 67.5% de los pacientes como no diabéticos frente al 32.5% clasificados como diabéticos, distribución consistente con la proporción original del *dataset* (65% no diabetes, 35% diabetes). De las ocho variables clínicas la más importante en la predicción del modelo es *Glucose*, con una importancia SHAP media de 0.112 y un valor de Random Forest de 0.324, y se define como la variable más influyente en la clasificación. Esto significa que niveles bajos de glucosa se asocian consistentemente con valores SHAP negativos (*efecto protector*), lo que reduce la probabilidad predicha de diagnóstico de diabetes y contrariamente niveles elevados de glucosa producen valores SHAP positivos (*efecto de riesgo*), e incrementa dicha probabilidad.

Por otra parte, el modelo LIME representado en la explicación de la figura 1a contiene una limitación metodológica relevante, la fidelidad promedio de las explicaciones locales fue de 0.41 por debajo del umbral aceptable de 0.80. No obstante, LIME demuestra ser eficaz como detector de variables relevantes a nivel individual e identifica las mismas *features* que SHAP en el 95% de los casos analizados según la concordancia por presencia. Esto indica que LIME identifica correctamente *cuáles* variables influyen en cada predicción individual, pero no cuantifica fielmente *cuánto* contribuye cada una, limitación inherente a la naturaleza de aproximación lineal *local* cuando se aplica a modelos de ensamble no lineales. La figura 1b presenta la concordancia (95%) y comparación de la presencia global de las variables en ambos modelos, con relación a la importancia normalizada para establecer diabetes en la población base.

1 La base de datos se puede encontrar en el siguiente enlace: https://www.kaggle.com/datasets/gzdekzlkaya/pima-indians-diabetes-dataset?select=pima_diabetes_data.csv

FIGURA 1. EXPLICACIÓN LIME (LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS).
a. GRAFICA DE EXPLICACIÓN DEL DATASET DE PIMA INDIANS DIABETES
b. GRÁFICA DE COMPARACIÓN ENTRE VALORES DE TÉCNICAS SHAP Y LIME



FUENTE: ELABORACIÓN PROPIA.

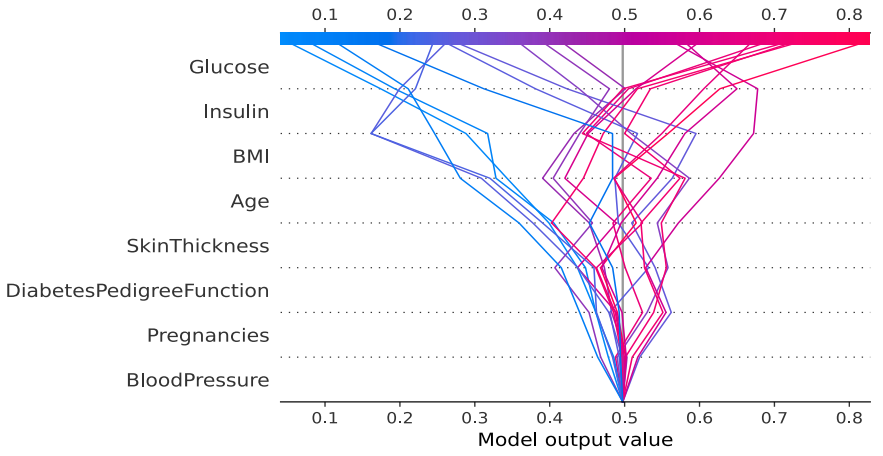
La Figura 2, inciso a) despliega la gráfica de decisión (*decision plot*) del modelo SHAP para las primeras 20 instancias del conjunto de prueba. Se observa que *Glucose* es la variable más influyente del modelo (*SHAP medio*=0.112), seguida de *Insulin* (0.076), *BMI* (0.062), *Age* (0.044) y *SkinThickness* (0.036). En la visualización los colores representan la dirección del efecto, los valores SHAP negativos (*azul*) indican contribución protectora contra el diagnóstico de diabetes, mientras que los valores SHAP positivos (*rojo*) indican contribución al riesgo de diabetes. Las predicciones que se aproximan a valores superiores a 0.50 se asocian con mayor probabilidad de diagnóstico positivo de diabetes, donde el valor base (*expected value*) del modelo es 0.497.

La Figura 2, inciso b) cascada (*waterfall plot*), presenta la descomposición aditiva de una predicción individual. Se toma como ejemplo la instancia 49 ubicada en la zona de incerti-

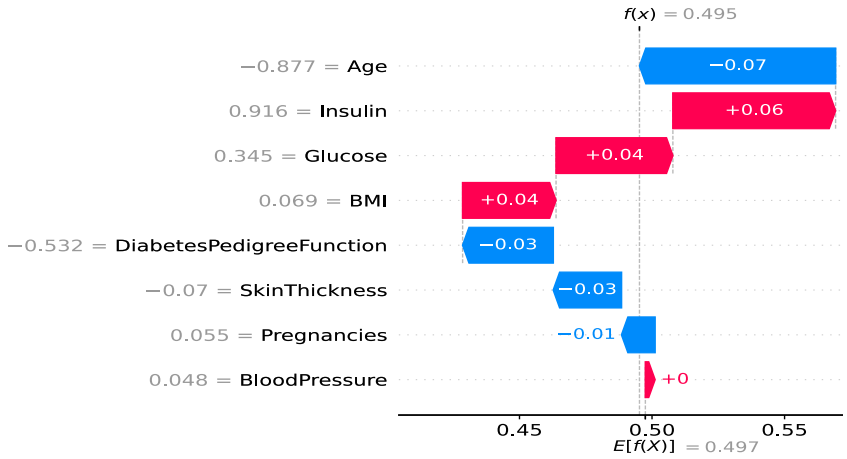
dumbre del modelo (*probabilidad predicha*=0.495), la predicción parte del valor base $E[f(X)]$=0.497 y las contribuciones de cada variable la desplazan hasta el valor final $f(x)$=0.495. En este análisis las variables más importantes son: *Age* con efecto protector (-0.074), *Insulin* con efecto de riesgo (+0.061), *Glucose* con efecto de riesgo (+0.044) y *BMI* con efecto de riesgo (+0.036). Para las instancias de alta confianza como la instancia 111 (*probabilidad*=0.855), la descomposición muestra que *Glucose* (+0.184), *Insulin* (+0.062), *Age* (+0.039) y *BMI* (+0.037) contribuyen conjuntamente al riesgo que eleva la predicción desde el valor base de 0.497 hasta 0.855.

La visualización permite evidenciar que los niveles de glucosa constituyen el factor predictivo dominante, seguidos de la insulina y el índice de masa corporal, resultados coherentes con la literatura epidemiológica sobre factores de riesgo de la enfermedad diabetes tipo 2. En este contexto la explicación visual proporcionada por SHAP es fundamental en los procesos interpretativos de la toma de decisiones, tanto algorítmicas como pragmáticas al ofrecer descomposiciones aditivas exactas (*RMSE*=0.00) con garantías axiomáticas de justicia distributiva basadas en los valores de Shapley (Jena, 2023; Rehman & Saba, 2025). LIME complementa esta interpretación al confirmar la identidad de las variables relevantes a nivel individual, con una concordancia del 95% aunque las limitaciones de fidelidad (0.41) indican que las atribuciones cuantitativas se fundamenten en SHAP, cuando se empleen modelos de ensamble.

Figura 2. Gráficos SHAP (SHapley Additive exPlanations).
a. Gráfico de decisión con valores SHAP en el tiempo (para series temporales) para el dataset de Pima Indians Diabetes
b. Gráfico de cascada. Ambas son explicaciones post-hoc.



a.



b.

FUENTE: ELABORACIÓN PROPIA.

Saarela & Podgorelec (2024) consideran que tanto SHAP como LIME tienen un uso predominante en los modelos XAI, con una preferencia por SHAP debido a la capacidad y estabilidad de las garantías matemáticas que los soportan. Estos métodos también se consideran basados en perturbaciones (Wang, 2024), y exploran el nivel de influencia de cada variable dentro del modelo conjunto, modifican valores y analizan los impactos en los resultados o salidas del modelo. Para estos autores el dominio de los estudios XAI corresponde a la aplicabilidad de estas ideas. Incluyen escenarios *ante-hoc* o intrínsecos de transparencia donde el modelo es la explicación, son inherentes y se basan en estructuras simples como modelos de regresión lineal y también se definen como *agnósticos*, frente a modelos a *post-hoc* diseñados para elucidar razonamientos después de una fase de entrenamiento. A esto se suman los modelos de explicaciones *locales* y *globales* (Dafali *et al.*, 2023), los primeros focalizan aspectos individuales y los segundos profundizan en cada modelo de modo general como un todo.

La división de las dos perspectivas: específica (*model-specific o intrínsecos*) (capas, pesos y gradientes) y agnóstica (*model-agnostic*) (solo entradas y resultados), se debe a que la primera comprende las características específicas del modelo y la segunda se fundamenta en técnicas basadas en perturbaciones, oclusiones y muestreo como mecanismos de extracción de información (Wang, 2024), aunque pueden usar aproximaciones de gradiente numérico sin acceder al modelo directamente. Esta distinción entre ambos enfoques específico y agnóstico no es solo una diferencia de implementación, se basa en modos de producción de conocimiento (Hsieh *et al.*, 2024). La tabla 1 presenta las principales definiciones de caracterización y delimitación identificadas en los modelos XAI.

Tabla 1. Principales aspectos de la inteligencia artificial explicable (XAI)

Aspecto	Inteligencia Artificial Explicable (XAI)
Definición Operativa	Las técnicas XAI se refieren a una serie de procedimientos y enfoques que facilitan a los usuarios comprender y depositar confianza en los resultados generados por los algoritmos de aprendizaje automático, es decir, modelos de *caja negra*. Nacen como técnicas de interpretabilidad de estos modelos, por ejemplo, LIME, SHAP, son métricas de evaluación que incluyen precisión, fidelidad y comprensión del usuario.
Métodos	**Modelos intrínsecamente interpretables**: Algoritmos diseñados para ser comprensibles por sí mismos, como regresiones lineales y árboles de decisión.
	Modelos de interpretación *post-hoc*: Técnicas que explican decisiones de modelos complejos después del entrenamiento y la ejecución.
	Visualización de datos: Herramientas que representan gráficamente el proceso de decisión de la IA para facilitar procesos de comprensión (PDP, ICE), anchors.
Técnicas XAI	**Técnicas de diagnóstico de modelos** (***post-hoc model-agnostic***): LIME, SHAP, kernelSHAP, treeSHAP, DeepSHAP, anchors, counterfactual explanations, partial dependence plots, individual conditional expectation plots, permutation feature importance, accumulated local effects, LORE. **Pós-hoc modelo específico (gradientes) (*post-hoc model-specific*)**: saliency maps, integrated gradients, Grad-CAM, smoothGrad, deepLIFT, layer-wise relevance propagation. **Basadas en conceptos**: TCAV, concept bottleneck models. **Basadas en ejemplos**: Prototipos y críticos, influential instances. **Destilación para explicabilidad**: destilación de modeles. **Explicación en lenguaje natural**: natural language explanations. **Ante-hoc integradas**: self-explaining neural networks, concept bottleneck models.
Modelos relacionados: de aprendizaje automático y estadístico	**Árboles de decisión**: CART, ID3, C4.5, C5.0, CHAID, conditional inference trees, oblique decision trees, random forest, gradient boosted trees, XGBoost, LightGBM, CatBoost, BART. **Modelos lineales**: regresión lineal simple, regresión lineal múltiple, regression logística binaria, regression logística multinomial, regresión logística ordinal, regresión probit, regresión de poisson y la regresión binomial negativa, regresión cuantílica, ridge regression, elastic net, linear discriminant analysis, perceptrón simple, SGD classifier. **Redes neuronales con mecanismos de atención**: Bahdanau Attention, Luong Attention, Transformer, Multi-Head Self-Attention, Cross-Attention, BERT, RoBERTa, ALBERT, DeBERTa, XLNet, GPT, GPT-2, GPT-3, GPT-4, T5,

	BART, PaLM, PaLM 2, Gemini, LLaMA, LLaMA 2, Claude, Mistral, Mixtral, Falcon, Vision Transformer, DETR, Swin Transformer, Whisper, Music Transformer, Longformer, BigBird, Flash Attention, Mixture of Experts. **Redes bayesianas**: red bayesiana, naive bayes, naive bayes gaussiano, naive bayes multinomial, naive bayes bernoulli, tree-augmented naive bayes, dynamic bayesian networks, hidden markov models, bayesian linear regression, bayesian logistic regression, gaussian processes, bayesian neural networks, latent dirichlet allocation, bayesian optimization, probabilistic graphical models, markov random fields, conditional random fields, variational autoencoders, bayesian nonparametrics, MCMC, Stan, PyMC, JAGS, BUGS. **Modelos de caja blanca:** regresión lineal, regresión logística, cart, C4.5, naive bayes, K-NN, tablas de decisión, generalized additive models, explainable boosting machines, neural additive models, bayesian rule lists, falling rule lists, rulefit, slim, scoring systems, programas lógicos inductivos, self-explaining neural networks, concept bottleneck models.
Objetivo principal	Transformar y mejorar los sistemas de IA de manera transparente, interpretable y confiable.
Propósitos	**Transparencia en decisiones de IA**: Permitir a los usuarios entender y confiar en las decisiones tomadas por sistemas de IA.
	Identificación y mitigación de sesgos: Detectar y corregir posibles prejuicios en los modelos de IA.
	Cumplimiento regulatorio: Asegurar que los sistemas de IA operen dentro de los marcos legales y éticos establecidos en cada dominio.
Propósitos	Explicaciones *post-hoc*, por ejemplo, LIME, SHAP. Modelos intrínsecamente interpretables, por ejemplo, árboles de decisión.
Inferencia	Inferencia causal en la generación de explicaciones para descubrir procesos de toma de decisiones de la IA. La inferencia es funcional y estadística sobre el comportamiento del modelo y los procesos de toma de decisiones. Posibles explicaciones *locales* frente a planteamientos *globales*.
Fuentes de datos	Conjuntos de datos de formación (*datsets*), resultados de modelos, atribuciones de características, y trabajo de campo.
Herramientas y encuadres	SHAP, LIME, interpretML, kit de herramientas de explicabilidad, tensorflow.
Éticas	Equidad y parcialidad en los modelos de IA, responsabilidad por las decisiones de IA.

FUENTE: ELABORACIÓN PROPIA A PARTIR DE GROBRÜGGE ET AL. (2024), HSIEH ET AL. (2024), KALASAMPATH ET AL. (2025), LUNDBERG & LEE (2018), MATHEW ET AL. (2025), Y WANG (2024).

La interpretabilidad como imperativo de los modelos XAI puede aplicarse a algunas de las líneas de investigación y aplicaciones en campos como las humanidades, ciencias sociales y modelos computacionales (Hofman *et al.*, 2021; Luján-Villar & Luján-Villar, 2019). Joyce *et al.* (2023) abordan la necesidad de claridad en la interpretabilidad de los modelos XAI en campos como la psiquiatría y la multiplicidad de variables que se asocian con síndromes, desenlaces, trastornos, signos y síntomas. Este aspecto complejiza los modelos predictivos a partir de relaciones probabilísticas entre sí, algo que se suma a etiologías posibles y determinantes sociopsicológicos de carácter multifactorial en múltiples trastornos. A nivel experimental estos enfoques metodológicos demuestran que el método de clasificación XGBoost y los análisis *post-hoc* de valores SHAP son eficientes en el tratamiento de los cambios de los estados de ánimo de una persona.

Los autores argumentan sobre la necesidad de delinear de manera específica y esclarecer los niveles de interpretabilidad de los modelos, por ejemplo, determinar la correspondencia entre las entradas y las características, someter a examen el rendimiento, y especialmente garantizar los criterios de interpretabilidad que cada proyecto exige. Este aspecto requiere esfuerzos decididos en la reflexividad científica y tecnológica de la relación entre humanos y máquinas, a propósito del uso de modelos predictivos en las ciencias sociales computacionales o complejas (Hofman *et al.*, 2021; Luján-Villar, 2018). Por lo menos tres aspectos son menester para establecer un nivel adecuado de interpretabilidad; validez, fiabilidad y comprensibilidad, esto incluye modelos diagnósticos y métricas performáticas o de rendimiento de los algoritmos implementados.

DISCUSIÓN

Modelos interpretables y técnicas XAI

En esta parte de la investigación se realiza una propuesta y discusión de los modelos interpretables y técnicas XAI de manera documentada, con la finalidad de aportar evidencia sobre la taxonomía de las técnicas XAI y profundizar la relación entre la IA explicable y la IA interpretable. En el mundo XAI la dependencia del modelo se refiere a cuánto *conoce* el método de explicación sobre la estructura interna del algoritmo que analiza. La tabla 2 y la Figura 3 resumen la taxonomía alternativa de las técnicas XAI disponibles, se analiza la tipología de cada proceso, el alcance en términos definibles de clasificación con relación a *cuándo* y *cómo* se genera la explicación de un algoritmo, y la dependencia de cada opción mediante categorías que

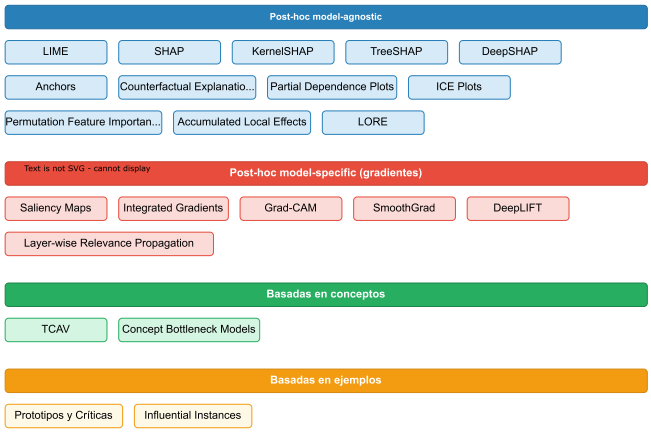
definen qué tan ligada está una técnica de explicación respecto a la arquitectura interna del modelo que intenta descifrar (Barredo *et al.*, 2020; Dib & Capus, 2025).

Tabla 2. Taxonomía de los XAI

Técnica	Tipo	Alcance	Dependencia del modelo
LIME	*Post-hoc*	Local	Agnóstico (generalista - versátil)
SHAP	*Post-hoc*	Local - global	Agnóstico (generalista - versátil)
Anchors	*Post-hoc*	Local	Agnóstico (generalista - versátil)
Contrafactuales	*Post-hoc*	Local	Agnóstico (generalista - versátil)
PDP	*Post-hoc*	Global	Agnóstico (generalista - versátil)
ICE	*Post-hoc*	Local	Agnóstico (generalista - versátil)
ALE	*Post-hoc*	Global	Agnóstico (generalista - versátil)
Permutation Importance	*Post-hoc*	Global	Agnóstico (generalista - versátil)
Saliency Maps	*Post-hoc*	Local	Específico (Redes Neuronales NN)
Integrated Gradients	*Post-hoc*	Local	Específico (Redes Neuronales NN NN)
Grad-CAM	*Post-hoc*	Local	Específico (CNN)
DeepLIFT	*Post-hoc*	Local	Específico (Redes neuronales NN)
LRP	*Post-hoc*	Local	Específico (Redes neuronales NN)
Atención	*Post-hoc - intrínseco*	Local	Específico (transformer, mapas de atención, relacional - contextual)
TCAV	*Post-hoc*	Global	Específico (Redes neuronales NN)
CBM	*Ante-hoc*	Global	Específico (técnico - matemático)
Árboles de decisión	*Ante-hoc*	Global	Intrínseco (simple - auto-explicativo)
GAMs/EBM	*Ante-hoc*	Global	Intrínseco (simple - auto-explicativo)
Rule Lists	*Ante-hoc*	Global	Intrínseco (simple - auto-explicativo)
Prototipos/Críticas	*Post-hoc*	Global	Agnóstico (generalista - versátil)
Influential Instances	*Post-hoc*	Local	Agnóstico (generalista - versátil)
Destilación	*Post-hoc*	Global	Agnóstico (generalista - versátil)
NLE	*Post-hoc*	Local	Específico (técnico - matemático)

Fuente: elaboración propia y Dib & Capus (2025).

FIGURA 3. TÉCNICAS DE INTELIGENCIA ARTIFICIAL EXPLICABLE (XAI)



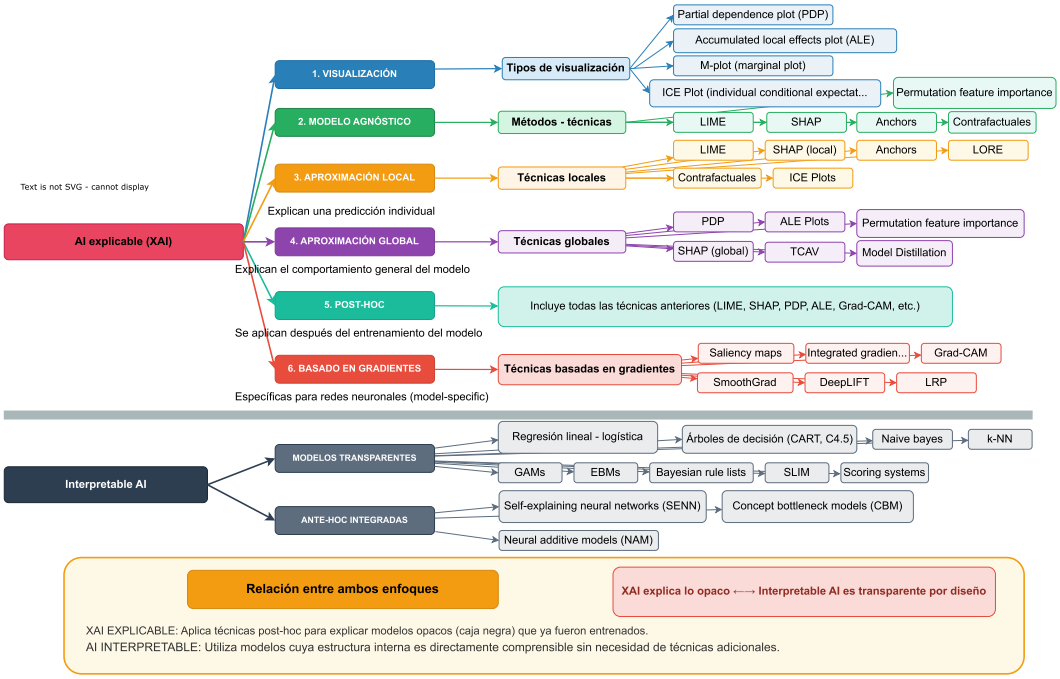
Resumen clasificatorio

Técnica	Tipo	Alcance	Dependencia
LIME	Post-hoc	Local	Agnostic
SHAP	Post-hoc	Local / Global	Agnostic
Anchors	Post-hoc	Local	Agnostic
Counterfactual Expl.	Post-hoc	Local	Agnostic
PDP	Post-hoc	Global	Agnostic
ICE Plots	Post-hoc	Local	Agnostic
Permutation F. Imp.	Post-hoc	Global	Agnostic
ALE Plots	Post-hoc	Global	Agnostic
LORE	Post-hoc	Local	Agnostic
Saliency Maps	Post-hoc	Local	Specific (NN)
Integrated Gradients	Post-hoc	Local	Specific (NN)
Grad-CAM	Post-hoc	Local	Specific (CNN)
SmoothGrad	Post-hoc	Local	Specific (NN)
DeepLIFT	Post-hoc	Local	Specific (NN)
LRP	Post-hoc	Local	Specific (NN)
TCAV	Post-hoc	Global	Specific (NN)
Concept Bottleneck	Ante-hoc	Global	Specific (NN)
Prototipos y Críticas	Post-hoc	Global	Agnostic
Influential Instances	Post-hoc	Local	Agnostic

FUENTE: ELABORACIÓN PROPIA.

La explicabilidad se centra en proporcionar inferencias interpretables por los humanos en términos de problemas; 1) para identificar decisiones del modelo a través de lenguaje natural, la provisión de visualizaciones o resultados relevantes; 2) explicar contrafactuales al evidenciar el razonamiento lógico y conocer las justificaciones que el modelo sostiene; y 3) este proceso puede incluir comparaciones para profundizar en el contexto de la modelización. En este punto se debe manifestar que las explicaciones generadas por técnicas *post-hoc* no son reveladoras de la verdad interna del modelo, proporcionan representaciones comprensibles de los problemas basados en aproximaciones estadísticas o funcionales para simplificar la comprensión del modelo. De manera complementaria la interpretabilidad se centra en la posibilidad que brinda a los investigadores de comprender los funcionamientos y prestaciones internas del modelo. Los modelos interpretables se basan en estructuras matemáticas o lógicas que ayudan la comprensión directa del mecanismo de decisión humana sin inferencias adicionales. Como se observa en la Figura 4 la interpretabilidad es inherente a estas arquitecturas por diseño, y se identifica cómo influyen las variables de entrada en la predicción (De *et al.*, 2025; Dib & Capus, 2025).

FIGURA 4. RELACIÓN ENTRE IA EXPLICABLE (XAI) E IA INTERPRETABLE



FUENTE: ELABORACIÓN PROPIA CON BASE EN DE ET AL. (2025).

La Tabla 3 resume los niveles de explicación en los modelos interpretables y técnicas XAI de manera diferenciada, mediante las siguientes dimensiones: fuente de interpretabilidad, naturaleza de la explicación, capacidad de modelado y dependencia. Este barrido identifica la diferencia de la arquitectura y la explicabilidad a nivel de modelo y técnica. A propósito de la compensación entre complejidad representacional e interpretabilidad directa, este equilibrio se basa en la transparencia de los modelos sin perder capacidad representacional en la toma de decisiones de alto riesgo para capturar relaciones altamente complejas.

Tabla 3. Niveles de explicación en modelos interpretables y técnicas XAI

Dimensión	Modelos interpretables	Técnicas XAI (*post-hoc*)
Fuente de interpretabilidad	La arquitectura del modelo mismo permite la comprensión directa de cómo se procesan las entradas.	Son algoritmos externos (LIME, SHAP, Grad-CAM, entre otros) que aproximan, descomponen o simulan el comportamiento del modelo original.
Naturaleza de la explicación	La lógica interna es explícita y transparente por diseño. El humano puede seguir el razonamiento paso a paso.	Se generan representaciones aproximadas localizadas o sintéticas que imitan o explican partes del modelo, pero no revelan la lógica completa si esta es compleja.
Capacidad de modelado	Son modelos *globales* por definición debido a la transparencia de la lógica interna que soportan, aunque presentan un espectro variable de capacidad representacional. Pueden capturar reglas discretas, sumas ponderadas, y árboles. Técnicas como GAMs, EBMs y Random Forests capturan no linealidad con rendimiento competitivo, y técnicas lineales como la regresión lineal y logística capturan exclusivamente relaciones lineales aditivas.	Estas técnicas no poseen capacidad propia de modelado y tienen como función generar explicaciones de modelos ya entrenados que serían opacos de otra forma. Permiten interpretar el comportamiento predictivo de modelos *globales* altamente complejos y no lineales, como redes neuronales profundas y ensambles de árboles. La explicación resultante es siempre una aproximación del modelo real, no el modelo *per se*. La fidelidad varía según la técnica y complejidad del modelo explicado.
Dependencia	No dependen de técnicas externas debido a que la interpretabilidad es una propiedad intrínseca del modelo.	Son metodologías por lo general independientes (agnósticas) aplicadas sobre cualquier modelo predictivo, aunque pueden no ser siempre independientes, por ejemplo, Grad-CAM o gradientes integrados.

Fuente: elaboración propia con base en De et al. (2025).

CONCLUSIONES

El campo XAI se encuentra en una tensión productiva entre la demanda de fidelidad explicativa: ¿Qué tan bien una explicación reproduce el comportamiento real del modelo?, ¿qué se evalúa en la comprensibilidad humana?, ¿los usuarios pueden entender la explicación generada? Hasta ahora difícilmente alguna técnica resuelve estas demandas simultáneamente de forma óptima. La elección de cada técnica debe ser contextualmente justificada según el problema investigativo, el dominio de aplicación, el tipo de modelo y los requisitos regulatorios**.** La confusión entre *modelo interpretable* y *técnica de explicación*, y los términos interpretabilidad, explicabilidad y transparencia es uno de los debates conceptuales más frecuentes en la literatura sobre XAI (Barredo *et al.*, 2020; Longo *et al.*, 2024). Hoy se sabe que la *predicción* es una prueba preliminar de la *explicación*, debido a que ambos aspectos de manera alineada reducen posibles ambigüedades interpretativas, un requisito epistémico en términos científicos (Saarela & Podgorelec, 2024).

La compensación o *trade-off* entre interpretabilidad y precisión es una guía importante en la investigación social. Un ejemplo práctico es el uso de redes neuronales para predecir comportamientos sistemáticos que pueden ser sociales, luego se pueden desentrañar las sensibilidades del modelo con *contrafactuales,* por ejemplo, *si x cambia* la predicción cambiaría. Este tipo de decisiones permite que el investigador pueda identificar finalmente, mecanismos reales o causales con el rigor necesario, una integración metodológica donde el modelo predictivo debe ser interpretable (Hofman *et al.*, 2021). Algunos desafíos epistémicos inherentes a la explicación de los sistemas complejos enfatizan sobre cómo el conocimiento teórico previo puede ser limitado o incompleto, una situación paradigmática necesaria en las investigaciones humanísticas y sociológicas contemporáneas. Uno de los principales aportes de las explicaciones interpretables a las ciencias sociales es la capacidad de simular mentalmente modelos y generar hipótesis plausibles para pruebas posteriores (Miller, 2019). Esto es especialmente relevante cuando los fenómenos sociales complejos presentan datos escasos o ruidosos.

En este contexto una distinción crítica para las humanidades se fundamenta en la interpretación cualitativa de resultados, lo que puede tener valor independiente o complementario del rendimiento cuantitativo final de una aplicación formal. Este razonamiento se basa en la experiencia subjetiva, los contextos situacionales, similitudes, diferencias, contrastes y, en síntesis, la consideración de matices cualitativos en la producción de sentido sobre fenómenos diversos que posibilitan establecer comparaciones, razonamientos y procedimientos exegéticos que no implican necesariamente precisión predictiva o demostrabilidad causal. La importancia de la explicabilidad e interpretabilidad no se basa en los formalismos que brindan claridad a procesos científicos que obtienen niveles de éxito predictivo en la modelación del mundo

social debido a que mantienen independencia lógica entre interpretabilidad, causalidad y predicción (Hofman *et al.*, 2021). Como se puede observar en el apartado de aplicaciones de este trabajo un nivel adecuado de interpretabilidad debe tener procesos validación, fiabilidad en la información y los datos y comprensibilidad en los resultados, lo que se complementa con modelos de diagnósticos y métricas performáticas o de rendimiento algorítmico.

Es importante manifestar que la explicabilidad es una condición necesaria en la modelación de la ciencia contemporánea, pero de manera aislada está muy lejos de ser suficiente para alcanzar justicia algorítmica *per se*. Un sistema puede ser perfectamente explicable y profundamente injusto en los resultados que provee. Las ciencias sociales y las humanidades no son un adorno ético para proyectos técnicos, mejor se deben comprender como la única forma de enunciar cuestionamientos éticos, políticos, culturales y asimetrías del poder que la naturaleza constitutiva de la técnica, no puede hacerse a sí misma (Luján-Villar, 2018). Cuando la responsabilidad se deposita en técnicas XAI se puede aspirar a tener la trazabilidad técnica del proceso, pero solamente las ciencias sociales y humanas pueden profundizar en los aspectos morales de la responsabilidad colectiva y política que tienen las aplicaciones algorítmicas. Este trabajo es un aporte en este sentido preciso.

El paradigma de la *realización* científica guía la gama de preguntas de investigación que conforman los diseños, modelos y programas investigativos. Se puede avanzar en formas de conocimiento fundamentado que soporten cuestiones pertinentes de la realidad del mundo, sin la necesidad de buscar verdades absolutas. Reconocer las epistemologías humanas posibilita identificar los tipos de razonamiento para considerar los niveles de veracidad con relación a diversos campos investigativos. Las metodologías científicas junto a técnicas y teorías buscan otorgar solidez a la imaginación investigativa, profundizar problemáticas complejas y prácticas de indagación alrededor de visiones sobre la búsqueda del conocimiento humano. Este paradigma científico busca ir más allá de descripciones y conjeturas, a través de espacios de diálogo continuo entre disciplinas, poblaciones y tipos de problemas investigativos.

REFERENCIAS

Aluvalu, R., Mehta, M., & Siarry, P. (2024). *Explainable AI in health informatics*. Springer Nature Singapore. https://doi.org/10.1007/978-981-97-3705-5

Barredo, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, 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

Chakraborty, D., Başağaoğlu, H., & Winterle, J. (2021). Interpretable vs. noninterpretable machine learning models for data-driven hydro-climatological process modeling. *Expert Systems with Applications*, *170*, 114498. https://doi.org/10.1016/j.eswa.2020.114498

Dafali, S. M., Kissi, M., & El Beggar, O. (2023). Comparative study between global and local explainable models. En Kissi, M. & El Beggar, O. (2023). *2023 14th International Conference on Intelligent Systems: Theories and Applications (SITA)* (pp. 1-8). IEEE. https://doi.org/10.1109/sita60746.2023.10373599

De, A., Saraf, S., Mishra, T.K., & Tripathy, B.K. (2024). Interpretation and visualization techniques in AI systems and applications. En Tripathy, B.K., & Seetha, H. (Eds.). (2024). *Explainable, interpretable, and transparent AI systems* (pp. 279-301). CRC Press. https://doi.org/10.1201/9781003442509

Dib, L., & Capus, L. (2025). Classifying XAI methods to resolve conceptual ambiguity. *Technologies*, *13*(9), 390. https://doi.org/10.3390/technologies13090390

Gaur, L., & Abraham, A. (Eds.). (2024). *Role of explainable artificial intelligence in E-Commerce*. Springer. https://doi.org/10.1007/978-3-031-55615-9

Grobrügge, A., Mishra, N., Jakubik, J., & Satzger, G. (2024). Explainability in AI-Based Applications – A Framework for Comparing Different Techniques. En Fettke, P., Weigand, H., Frank, U., Huemer, C., & Schnellmann, M. (2024). *2024 26th International Conference on Business Informatics (CBI)*, Vienna, Austria, 2024 (pp. 70-79). https://doi.org/10.1109/cbi62504.2024.00018

Hofman, J.M., Watts, D.J., Athey, S., Garip, F., Griffiths, T.L., Kleinberg, J., ... & Yarkoni, T. (2021). Integrating explanation and prediction in computational social science. *Nature*, *595*(7866), 181–188. https://doi.org/10.1038/s41586-021-03659-0

Hsieh, W., Bi, Z., Jiang, C., Liu, J., Peng, B., Zhang, S., ... & Liu, M. (2024). A comprehensive guide to explainable AI: from classical models to LLMs. *arXiv preprint. arXiv:2412.00800*. https://doi.org/10.48550/arXiv.2412.00800

Huang, F., Jiang, S., Li, L., Zhang, Y., Zhang, Y., Zhang, R., Li, Q., Li, D., Shangguan, W., & Dai, Y. (2024). Applications of explainable artificial intelligence in earth system science. *arXiv preprint. arXiv:2406*.11882. https://doi.org/10.48550/arXiv.2406.11882

Jabareen, Y. (2009). Building a conceptual framework: Philosophy, definitions, and procedure. *International Journal of Qualitative Methods, 8*(4), 49–62. https://doi.org/10.1177/160940690900800406

Jena, K. (2023). Innovations in Explainable AI: Bridging the Gap Between Complexity and Understanding. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *9*(6), 118-125. https://doi.org/10.32628/cseit2390613

Joyce, D.W., Kormilitzin, A., Smith, K.A., & Cipriani, A. (2023). Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *npj Digital Medicine, 6*(6). https://doi.org/10.1038/s41746-023-00751-9

Juarez, J.M., Bielza, C., Fernandez-Llatas, C., Johnson, O., Kocbek, P., Larrañaga, P., Johnson, O., Štiglic, G., Martin, N., Munoz-Gama, G., Sepulveda, M., & Vellido, F. (2024). *Explainable artificial intelligence and process mining applications for healthcare*. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-54303-6

Kalasampath, K., Spoorthi, K. N., Sajeev, S., Kuppa, S. S., Ajay, K., & Angulakshmi, M. (2025). A Literature review on applications of explainable artificial intelligence (XAI). *IEEE Access*, *13*, 41111-41140. https://doi.org/10.1109/ACCESS.2025.3546681

Khamparia, A., & Gupta, D. (Eds.). (2025). *Explainable artificial intelligence for biomedical and healthcare applications*. CRC Press. https://doi.org/10.1201/9781003220107

Kumar, A., Kumar, T.A., Das, P., Sharma, C., & Kumar, A. (2025). *Explainable artificial intelligence in the healthcare industry*. Scrivener Publishing, Wiley. https://doi.org/10.1002/9781394249312

Liu, Q., Gui, D., Zhang, L., Niu, J., Dai, H., Wei, G., & Hu, B.X., (2022). Simulation of regional groundwater levels in arid regions using interpretable machine learning models. *Science of The Total Environment*, *831*, 154902. https://doi.org/10.1016/j.scitotenv.2022.154902

Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Del Ser, J., ... & Stumpf, S. (2024). Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, *106*, 102301. https://doi.org/10.1016/j.inffus.2024.102301

Luján-Villar, J.D. (2018). Ciencias sociales y sostenibilidad: tecnologías de investigación social aplicadas a lo urbano y lo rural. *Revista de Antropología y Sociología: VIRAJES*, *20*(1), 61-81. https://doi.org/10.17151/rasv.2018.20.1.4

Luján-Villar, J.D., & Luján-Villar, R.C. (2019). Complejidad y ciencia de sistemas: aproximación a su impacto actual en el mundo. *CienciAmérica*, *8*(2), 103-122. https://doi.org/10.33210/ca.v8i2.234

Lundberg, S.M., & Lee, S.-I. (2018). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NeurIPS), 31*, 4765–4774. https://doi.org/10.48550/arXiv.1705.07874

Malviya, R., & Sundram, S. (2025). *Explainable and responsible artificial intelligence in healthcare*. Scrivener Publishing, Wiley. https://doi.org/10.1002/9781394302444

Mathew, D.E., Ebem, D.U., Ikegwu, A.C. Eberechukwu, P., & Dibiaezue, N.F. (2025). Recent emerging techniques in explainable artificial intelligence to enhance the interpretable and understanding of AI models for human. *Neural Processing Letters, 57*(16), 1-32. https://doi.org/10.1007/s11063-025-11732-2

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*. https://christophm.github.io/interpretable-ml-book/

Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B., (2019). Interpretable machine learning: definitions, methods, and applications. *Proceedings of the National Academy of Sciences of the United States of America*, *116*, 22071–22080. https://doi.org/10.1073/pnas.1900654116

Pretolesi, D., Stanzani, I., Ravera, S., Vian, A., Barla, A. (2025). Artificial intelligence and network science as tools to illustrate academic research evolution in interdisciplinary fields: The case of Italian design. *PLOS ONE, 20*(1), e0315216. https://doi.org/10.1371/journal.pone.0315216

Rehman, A., & Saba, T. (Eds.). (2025). *Explainable artificial intelligence in medical imaging*. Auerbach Publications, CRC Press. https://doi.org/10.1201/9781032626345

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, *16*, 1-85. https://doi.org/10.1214/21-SS133

Saarela, M., & Podgorelec, V. (2024). Recent applications of explainable AI (XAI): A systematic literature review. *Applied Sciences*, *14*(19), 8884. https://doi.org/10.3390/app14198884

Wang, Y. (2024). A comparative analysis of model agnostic techniques for explainable artificial intelligence. *Research Reports on Computer Science, 3*(2), 25-33. https://doi.org/10.37256/rrcs.3220244750

Whittemore, R., & Knafl, K. (2005). The integrative review: Updated methodology. *Journal of Advanced Nursing*, *52*(5), 546–553. https://doi.org/10.1111/j.1365-2648.2005.03621.x

Zhang, Q., Ding, K., Lv, T., Wang, X., Yin, Q., Zhang, Y., ... & Chen, H. (2025). Scientific large language models: A survey on biological & chemical domains. *ACM Computing Surveys*, *57*(6), 1-38. https://doi.org/10.1145/3715318