Revista Facultad de Ingeniería (Rev. Fac. Ing.) Vol. 33, No. 70.
L-ISSN: 0121-1129, e-ISSN: 2357-5328.

DOI: 10.19053/01211129.v33.n70.2024.18078

REVISIÓN SISTEMÁTICA DE LAS HERRAMIENTAS DE INTELIGENCIA ARTIFICIAL EXPLICABLE USADAS EN MÉTODOS DE ENSAMBLE

Explainable Artificial Intelligence tools used in ensemble methods: A systematic literature review

Fabian-Camilo Martínez-Silva
Universidad del Cauca (Popayán, Colombia).
fcmartinez@unicauca.edu.co

Carlos-Alberto Cobos-Lozada
Universidad del Cauca (Popayán, Colombia).
ccobos@unicauca.edu.co

Fecha de recibido: 27-08-2024
Fecha de aceptado: 2-12-2024

RESUMEN

Los modelos de aprendizaje automático se utilizan cada vez más por su alto rendimiento predictivo. Sin embargo, en ámbitos críticos como la salud, la seguridad y la defensa y las finanzas existe una necesidad urgente de modelos que también sean explicables. Por lo general, los modelos más complejos (como las redes neuronales profundas y los ensambles) obtienen los mejores resultados en problemas de gran envergadura. Pese a esto, su falta de transparencia limita su aplicación en áreas donde se requiere comprender el proceso de toma de decisiones. Teniendo en cuenta que los modelos de ensamble obtienen los mejores rendimientos en diversas aplicaciones —como se aprecia en las competencias de Kaggle, en particular XGBoost—, en este artículo se presenta una revisión sistemática de los artículos publicados entre el 2019 y el 2024, relacionados con el uso de herramientas de Explicabilidad de la Inteligencia Artificial (XAI) en modelos de ensamble. La metodología seguida para la revisión se basó en las directrices propuestas por Kitchenham et al. que consideran la planeación y la ejecución de la revisión junto con el reporte de los resultados obtenidos. Los resultados de la investigación permiten comprender los beneficios y desafíos del uso de XAI en los ensambles para apoyar la toma de decisiones y contribuir al derecho social de la explicación de estas. Además, al identificar las áreas en las que se está investigando y los contextos donde más se aplica, se visibilizan otras áreas de oportunidad para futuras investigaciones. Se concluye que existen herramientas y enfoques prometedores que permiten una mejor comprensión de la lógica de los modelos de ensamble y una mayor transparencia en los resultados, y se identifican áreas de mejora y la necesidad de continuar investigando para abordar los desafíos asociados con la explicabilidad de los modelos de ensamble.

Palabras clave: aprendizaje automático interpretable; Inteligencia Artificial Explicable (XAI); modelos de ensamble; técnicas de explicabilidad; XgBoost.

ABSTRACT

Machine Learning models are increasingly used due to their high predictive performance. However, in critical areas such as healthcare, security and defense, and finance, there is an urgent need for models that are also explainable. Generally,

Cómo citar este artículo: F. C. Martínez-Silva, C. A. Cobos-Lozada, "Revisión sistemática de las herramientas de inteligencia artificial explicable usadas en métodos de ensamble". *Revista Facultad de Ingeniería*, vol. 33, no. 70, e18078, 2024. https://doi.org/10.19053/01211129.v33.n70.2024.18078

more complex models —such as Deep Neural Networks and ensemble methods— achieve the best results in high-complexity problems; nevertheless, their lack of transparency limits their application in areas that require understanding the decision-making process. Given that ensemble models achieve the best performance in various applications (as seen in Kaggle competitions, particularly XGBoost), this paper presents a systematic literature review of articles published between 2019 and 2024 on the use of Explainable Artificial Intelligence (XAI) tools in ensemble models. The methodology is based on the guidelines proposed by Kitchenham and others, which include planning and execution of the review along with reporting the results obtained. Research results shed light on the benefits and challenges of using XAI in ensemble models to support decision-making and contribute to the social right of explaining such decisions. In addition to identifying research areas and contexts where these models are most applied, opportunities for future research are highlighted. It concludes that there are promising tools and approaches that enable a better understanding of the logic behind these models and greater transparency in results; it also identifies areas for improvement and the need to continue researching to address the challenges associated with the explainability of ensemble models

Keywords: ensemble models; Explainable Artificial Intelligence (XAI); explainability techniques; interpretable machine learning; XgBoost.

1. INTRODUCCIÓN

La inteligencia artificial (IA) ha experimentado un progreso significativo y continuo en la última década, lo que ha resultado en una mayor adopción de sus algoritmos, por ejemplo, el uso de algoritmos de aprendizaje profundo (Deep Learning - DL) y de aprendizaje automático (Machine Learning - ML) para resolver muchos problemas, incluso aquellos que eran difíciles de solucionar en el pasado [1]. Sin embargo, estos logros sobresalientes van acompañados de modelos con una mayor complejidad que se clasifican como de caja negra, ya que carecen de transparencia. Por lo tanto, se hace necesario encontrar soluciones que puedan contribuir a abordar este desafío (lograr calidad siendo transparente) para expandir el uso de los sistemas basados en IA en dominios críticos y sensibles como atención médica, aplicación de la ley y seguridad y defensa nacional.

La inteligencia artificial explicable (XAI) se ha propuesto como un área temática y de investigación que busca avanzar hacia una IA más transparente para evitar que se limite su adopción en dominios críticos. Los modelos de aprendizaje automático interpretables se definen como aquellos que se visualizan o explican claramente utilizando textos o gráficos y pueden ser fácilmente comprendidos por diversos tipos de usuarios [2].

Un árbol de decisión es un ejemplo de un modelo interpretable ampliamente utilizado. Por su parte, los bosques de decisión (un tipo de ensamble basado en *bagging*, *boosting* o *stacking*), como Random Forest o XGBoost, que combinan varios árboles de decisión para proporcionar un único resultado en tareas supervisadas de aprendizaje automático, son modelos no interpretables. Estos han demostrado recientemente ser altamente efectivos en numerosas tareas, pero todos se consideran modelos de caja negra [3].

El presente artículo presenta una revisión sistemática sobre las técnicas, métodos, o algoritmos XAI aplicados a modelos de ensamble. Se encuentra dividido en cuatro secciones. La primera corresponde a esta introducción. La segunda describe la metodología utilizada para desarrollar la revisión y los criterios que se tuvieron en cuenta para la selección de los artículos. La tercera muestra los resultados obtenidos tras realizar la búsqueda en las fuentes definidas y analizar los artículos seleccionados, destacando los beneficios y puntos a mejorar en la aplicación de XAI. Finalmente, se presentan las conclusiones y algunas ideas de trabajos futuro a desarrollar en el área.

2. METODOLOGÍA

La metodología se basó en las directrices presentadas por Kitchenham *et al*. [4] y constó de tres fases principales: la planeación de la revisión, su ejecución o realización y el reporte de los resultados obtenidos. La fase de planeación tuvo como objetivo definir el protocolo de la revisión que contempló la definición de: 1) objetivo y preguntas de investigación (Tabla 1); 2) cadenas de búsqueda; 3) fuentes de información (Scopus y Web of Science); 4) criterios de inclusión y exclusión de los artículos encontrados, basados en la lectura de los títulos, resúmenes, introducción, conclusiones y palabras clave de los artículos encontrados (Tabla 2); y 5) criterios evaluación de calidad de los artículos que fueron finalmente seleccionados, considerando su revisión completa .

En la fase de ejecución, se ingresó a las fuentes de información definidas y se ejecutaron las cadenas de búsqueda usando la sintaxis apropiada para cada una de ellas, los resultados se exportaron y se gestionaron en *Parsifal*. En la fase de reporte de los resultados se describieron los resultados de las dos fases anteriores, presentando de manera detallada los hallazgos obtenidos.

Tabla 1. Preguntas de investigación

Preguntas de investigación	Motivación
¿Qué innovaciones con enfoque XAI se han usado en modelos de aprendizaje automático basados en ensambles (*bagging*, *boosting*) para problemas de regresión y clasificación y cuáles han sido las ventajas y desventajas que se han observado en su uso?	Identificar cómo se ha aplicado XAI en diferentes modelos de ensamble (*bagging*, *boosting*) de aprendizaje automático y determinar los beneficios o innovaciones relevantes.
¿Qué trabajos previos transformaron un bosque de árboles tipo *bagging* o *boosting* en un solo árbol de decisión?	Encontrar qué soluciones se han propuesto para convertir una caja negra de un ensamble de árboles en una caja de cristal a través de un árbol de decisión y cómo lo hicieron.

Tabla 2. Criterios de inclusión y exclusión

Tipo de criterio	Criterio	Tipo
Lenguaje	Documento escrito en inglés o español.	Inclusión
Tipo de fuente	Documento del tipo libro, artículo o acta de conferencia.	
Accesibilidad	Documento de acceso libre o al que se logre acceder a partir de los convenios que tiene la Universidad del Cauca.	
Relevancia con las preguntas de investigación	El título, el resumen, la introducción, las conclusiones y las palabras clave del documento reflejan relevancia con, por lo menos, una pregunta de investigación de la revisión.	
Lenguaje	Documento escrito en un lenguaje diferente al inglés o al español.	Exclusión
Tipo de fuente	Documento que no corresponde al tipo libro, artículo o acta de conferencia.	
Accesibilidad	Documento no accesible.	
Relevancia con las preguntas de investigación	El título, el resumen, la introducción, las conclusiones y las palabras clave del documento no reflejan relevancia con, por lo menos, una pregunta de investigación de la revisión.	

La Figura 1 resume el proceso de búsqueda y selección de los estudios primarios relevantes para la presente revisión.

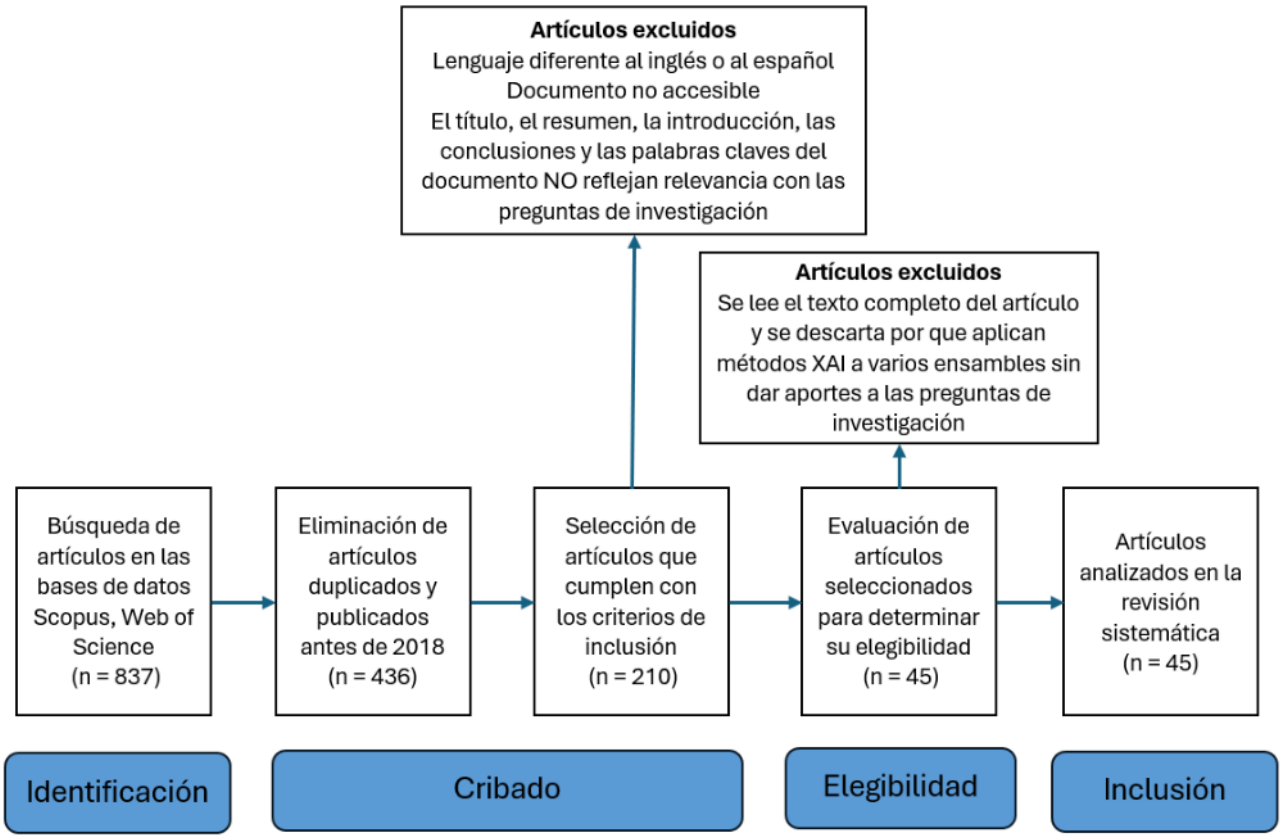

Figura 1. *Diagrama de flujo de la estrategia de búsqueda y selección de artículos*

A. Cadena de búsqueda y fuentes de información

(XAI OR "Explainable Artificial Intelligence" OR "Explainable machine learning" OR "Interpretable Artificial Intelligence" OR "Interpretable machine learning" OR "Interpretable AI") AND (ensemble OR bagging OR boosting OR stacking OR blending) AND (adaboost OR XGBoost OR "random forest" OR extratrees OR gradient OR lightgbm OR histgradientboosting OR "Mixture of Experts") AND (Publication Year BETWEEN 2019 AND 2024).

3. RESULTADOS

La búsqueda fue realizada el 25 de junio de 2024 y se encontraron en total 837 documentos. En la Tabla 4, se muestra el número de documentos recuperados y seleccionados por fuente, luego de aplicar los criterios de inclusión y exclusión (Tabla 3).

Tabla 3. Criterios de inclusión y exclusión

Criterio	Descripción	Puntuación		
		-1	**0**	**1**
C1	El estudio presenta una descripción detallada de los algoritmos y métodos utilizados.	No	Parcialmente	Sí
C2	El estudio presenta claramente y en detalle los resultados obtenidos.	No	Parcialmente	Sí
C3	El estudio ha sido publicado en una revista relevante.	No	JCR, Q3 o Q4	JCR, Q1 o Q2
C4	El estudio ha sido citado por otros autores	Una o menos citaciones promedio anuales a partir del año siguiente a su publicación.	Más de una y menos de tres citaciones promedio anuales a partir del año siguiente a su publicación.	Tres o más citaciones promedio anuales desde el año siguiente a su publicación o publicado en 2024.

Tabla 4. Número de resultados encontrados

Base de datos	Número de artículos encontrados	Seleccionados
Scopus	465	23
Web of Science	372	22
Total	**837**	**45**

La tendencia creciente en el número de publicaciones y el hecho de que la mayor cantidad de las seleccionadas fuera del año 2023 evidencia que el tema es de gran relevancia en la actualidad.

En los resultados de la evaluación de la calidad de los estudios seleccionados se pudo observar que 38 artículos recibieron la calificación máxima; cuatro recibieron una calificación de tres, que es buena (siendo castigados por tener un número promedio de dos citas al año); y tres recibieron una calificación de dos que es aceptable (siendo castigados por tener un número promedio de citas al año igual o menor a uno). Asimismo, el 100 % de los estudios realizó una descripción detallada de los algoritmos y métodos usados (C1), mostró los resultados de manera clara (C2), había sido publicados en revistas relevantes (Q1 o Q2) y, finalmente, el 85 % de los estudios obtuvieron tres o más citaciones promedio al año desde el año siguiente a su publicación o fueron publicados en 2024.

A continuación, se presenta el análisis de los artículos seleccionados siguiendo el tipo de método o técnica XAI utilizada, a saber: modelos agnósticos, basados en modelos, usando modelos locales, usando modelos globales o usando modelos sustitutos (*black box to glass box*).

A. XAI aplicado a modelos de ensamble basado en modelos agnósticos

Los modelos agnósticos son enfoques que se centran en la interpretación de modelos de aprendizaje automático sin considerar la arquitectura o el algoritmo de IA subyacente [5]; y pueden generar interpretaciones más comprensibles para los no expertos porque no requieren un conocimiento técnico del modelo [6]. Estas técnicas se basan en la salida del modelo y buscan proporcionar explicaciones generales aplicables a diferentes tipos de modelos.

Los artículos seleccionados proporcionaron una visión general de las técnicas existentes y su aplicabilidad centrándose en los valores Shapley Additive Explanations (SHAP) [7], que permiten

una interpretación global y local del ensamble al medir la contribución de cada característica en las predicciones del modelo; mientras que las técnicas XAI agnósticas del modelo provocan pérdida de información y solo dan una reinterpretación de la contribución de las características.

Los métodos de explicación *post hoc*, tanto los específicos del modelo como los agnósticos, actualmente son la corriente principal de investigación. En particular, la interpretación visual, la importancia de las características, así como la reducción del modelo son los que más comúnmente se utilizan para la explicación *post hoc* [8]. A su vez, dado que los métodos agnósticos se basan en el modelado *post hoc* de funciones arbitrarias, pueden ser ineficientes y sesgados debido al muestreo ponderado que realizan.

Para algunos autores [5], los métodos XAI actualmente disponibles para explicar GBT son los agnósticos. Estos incluyen explicaciones independientes del modelo localmente interpretables (LIME), aditivas de Shapley (SHAP) y basadas en reglas locales (LORE) y anclajes.

Se dice que estos métodos de propósito general pueden usarse para explicar cualquier modelo y esta flexibilidad se logra mediante el uso de un conjunto de datos sintéticos para sondear el modelo de caja negra de referencia e inferir una relación entre sus entradas y salidas [5]. Sin embargo, están en desventaja porque no hay introspección del modelo de referencia o distribución del objetivo que se considera esencial para explicaciones confiables. Además, las explicaciones pueden exhibir varianza debido a la generación no determinista de datos y la alta varianza da como resultado explicaciones diferentes para casos similares, lo cual genera una gran desconfianza en dichas explicaciones.

B. XAI aplicado a modelos de ensamble basado en modelos específicos

Un método de explicabilidad específico del modelo o intrínseco solo es aplicable a un tipo particular de modelo. Estos métodos utilizan las características y estructuras internas del modelo de aprendizaje automático, en lugar de depender únicamente de los datos de entrada y salida para proporcionar interpretaciones sobre cómo se realizan las predicciones [9].

Las técnicas más comunes de explicación basadas en modelos son la importancia de las características, las reglas de decisión y la visualización de modelos. La primera evalúa la significancia relativa de cada característica en el modelo teniendo en cuenta medidas como la importancia de Gini en árboles de decisión e identificando qué características tienen más influencia en las predicciones. Las reglas de decisión se utilizan para traducir la lógica interna del modelo en reglas comprensibles para los humanos [10] y describir cómo se toman las decisiones basadas en los valores de las características.

En la visualización de modelos se representan gráficamente la estructura y el funcionamiento interno de estos y puede incluir visualizaciones de árboles de decisión, diagramas de flujo o representaciones gráficas de las relaciones entre las características y las predicciones. Sin embargo, las explicaciones pueden variar según el tipo de modelo utilizado y no ser aplicables a todos los algoritmos de aprendizaje automático [11]. Además, diferentes métodos o implementaciones de aprendizaje en conjunto pueden conducir a diferentes características para la interpretación.

Un enfoque notable fue el desarrollado por algunos autores [12], donde se presentó un método para transformar un bosque de decisión generado por XGBoost en un único árbol de decisión interpretable. Este método permite simplificar la complejidad del modelo al crear un árbol que mantiene gran parte del rendimiento predictivo del bosque original, pero con una estructura mucho más comprensible para los usuarios finales.

Por su parte, otros autores [13] se propusieron un procedimiento para construir un árbol de decisión que aproximó el rendimiento de modelos de ensamble complejos como XGBoost. Su método se enfocó en la estabilidad y potencia de aproximación, proporcionando una herramienta valiosa para interpretar y simplificar patrones de predicción. Otro estudio [14] abordó la problemática de la interpretabilidad mediante la creación de un árbol de decisión único que se aproximó a un conjunto de árboles generado por XGBoost, utilizando las distribuciones de clases predichas por el conjunto. Finalmente, en otro caso [15], se introdujo un método para interpretar modelos de caja negra mediante la construcción de un árbol de decisión basado en la contribución de las variables de entrada a las predicciones facilitando la comprensión del modelo y el descubrimiento de nuevos conocimientos.

C. XAI aplicado a modelos de ensamble con modelos locales

La interpretabilidad local se logra cuando es posible comprender la lógica de una sola predicción (o de un conjunto de estas) sin comprender necesariamente toda la estructura del modelo. LIME manifiesta este enfoque aproximando el comportamiento del modelo en torno a predicciones específicas mediante un modelo interpretable más simple, permitiendo comprender decisiones individuales de manera efectiva [17].

Los métodos de explicación local tienen como objetivo obtener la influencia de las características de entrada en el resultado de la decisión de un caso específico de entrada. Los enfoques actuales de explicación local incluyen: 1) informar la ruta de la decisión, 2) aplicar varios métodos independientes del modelo, lo que requiere su ejecución repetidamente para cada predicción y 3) asignar crédito a cada característica de entrada [16].

Las técnicas de explicación local utilizadas en modelos de ensamble, reportadas en los artículos primarios seleccionados, incluyeron el cálculo de la importancia de cada característica en la predicción de una instancia específica, mediante el análisis del cambio en el valor del gradiente del modelo al perturbar el valor de dicha característica [10]. Esta técnica ofrece una medida del impacto significativo de las características en predicciones individuales [16].

Asimismo, el método SHAP se utilizó para asignar contribuciones individuales a cada característica basándose en la teoría de juegos y empleando los valores de Shapley para determinar la contribución justa de cada una en la predicción de una instancia específica o un conjunto de instancias (explicaciones locales) o en explicaciones globales.

Las explicaciones locales y la generación de ejemplos se consideran pertenecientes al enfoque de simplificación [17]. La técnica de generación de ejemplos extrae muestras de datos asociadas con los resultados de un modelo específico, permitiendo a los usuarios comprender mejor dicho modelo. Estudios como los presentados por algunos autores [18-20] abordaron el problema de la explicabilidad de los modelos combinando explicaciones locales para construir una comprensión global.

En un caso [18], usaron LIME para explicar modelos complejos como los ensambles, donde la explicación directa de la estructura completa del modelo es intrincada. En otro [19], emplearon SHAP para proporcionar tanto explicaciones locales como globales del modelo. También [21], se exploró el uso de LIME y SHAP discutiendo tanto sus aplicaciones como sus limitaciones y ofreciendo una visión integral del panorama actual en inteligencia artificial explicable (XA). En otro caso [22], propusieron un marco de evaluación para la interpretabilidad de los modelos de IA, subrayando la importancia de combinar explicaciones locales y globales para obtener una comprensión más completa del comportamiento del modelo.

Además, en otro estudio [23], se introdujeron técnicas de explicación contrafactual que muestran cómo los cambios en las características de entrada pueden alterar las predicciones del modelo proporcionando una perspectiva distinta sobre cómo se pueden interpretar las decisiones del modelo en situaciones específicas.

D. XAI aplicado a modelos de ensamble en modelos globales

Un modelo globalmente interpretable debe permitir una comprensión completa de su funcionamiento interno revelando no solo las características clave que influencian las predicciones sino también las interacciones y relaciones entre ellas, permitiendo la comprensión del razonamiento que conduce a todos sus resultados posibles [24-25]. Ejemplos de modelos interpretables globalmente son los árboles de decisión, las reglas de decisión y los modelos lineales. Algunos autores [26-27] presentaron métodos para generar reglas y árboles de decisión a partir de redes neuronales profundas preentrenadas y otros modelos de caja negra.

Las técnicas de explicación global utilizadas en ensambles se basan en medir la importancia relativa de cada característica con base en el cálculo de la ganancia de información obtenida por cada característica al dividir el conjunto de datos durante el entrenamiento del modelo, generando una clasificación de las características según su importancia en la toma de decisiones. Para este cálculo, se permutan aleatoriamente los valores de cada característica y se estima la disminución resultante en la precisión del modelo [28]. Esta técnica, aunque poderosa, puede ser computacionalmente costosa y sensible a la correlación entre características.

Los valores de SHAP también se pueden utilizar para evaluar la importancia global de cada característica en los modelos de ensamble [29-30], proporcionando una visión general de cómo cada característica contribuye al modelo y cómo se relacionan entre sí ayudando a identificar características irrelevantes o sesgadas.

Los Gráficos de Dependencia Parcial (Partial Dependence Plot - PDP) son una técnica que permite visualizar el efecto marginal de una característica sobre la predicción ofreciendo una perspectiva más detallada y permitiendo la identificación de relaciones no lineales y no monótonas entre las variables [31].

Algunos autores [32] propusieron modificaciones específicas en algoritmos de ensambles para mejorar su interpretabilidad global que consisten en limitar la profundidad de los árboles a uno, creando un modelo similar a un Modelo Aditivo Generalizado (GAM) donde cada árbol representa el efecto marginal de una característica individual.

Este enfoque simplifica la interpretación al evitar interacciones complejas entre características, pero, en general, sacrifica la capacidad del modelo para capturar dichas interacciones. Además, está basado en un conjunto de máquinas de aumento de gradiente (GBM) donde cada GBM se entrena en una única característica, por lo que el modelo global se construye como una suma ponderada de estos GBM individuales, facilitando tanto la interpretación global como local, minimiza el riesgo de sobreajuste y permite mantener un buen rendimiento.

En otro caso [33], introdujeron ExMatrix, una matriz que visualmente favorece la interpretación global y local de clasificadores basados en bosques aleatorios y permite a los usuarios identificar rápidamente patrones globales y relaciones entre características, así como examinar casos específicos para comprender las razones detrás de las predicciones individuales. Las filas representan reglas, las

columnas características y las celdas contienen predicados de reglas. Esta representación permite la visualización de un gran número de reglas de manera escalable.

E. XAI aplicada a modelos de ensamble con modelos sustitutos

El concepto de *black box to glass box* (de caja negra a caja de cristal) se refiere al proceso de convertir un modelo de aprendizaje automático opaco (caja negra) en un modelo interpretable y comprensible (caja de cristal), lo que también se conoce como un modelo sustituto [34]. En el caso de XGBoost, que es un modelo de ensamble tipo *boosting* basado en árboles, se introdujeron técnicas para mejorar su transparencia e interpretabilidad [35-36] que incluyeron: 1) las reglas de decisión, 2) la visualización de los árboles, 3) la simplificación del modelo y 4) la representación de las características.

Un enfoque destacado fue el propuesto por algunos autores [12, 37] que transformó un modelo de Gradient Boosting Decision Tree (GBDT), como XGBoost, en un único árbol de decisión interpretable. Este método implicó tres etapas clave: 1) la poda del conjunto para reducir el número de hojas, 2) la extracción de reglas conjuntivas del conjunto podado y 3) la organización de estas reglas en una estructura de árbol de decisión jerárquico. La configuración de la profundidad del árbol permitió balancear la precisión y la interpretabilidad, adaptándose a las necesidades del usuario.

Sin embargo, este método presenta varias desventajas. La principal es la posible pérdida de precisión predictiva en situaciones donde el modelo original es altamente complejo, ya que la simplificación de múltiples árboles en uno solo puede omitir interacciones importantes. La dependencia en la calidad de la poda y el riesgo de sobreajuste son preocupaciones adicionales, así como la complejidad computacional y la necesidad de una configuración experta para optimizar el árbol [38].

La visualización de los árboles de decisión utilizados en XGBoost puede ayudar a comprender la estructura del modelo y cómo se realiza la predicción [39]. Se pueden mostrar gráficamente los nodos, las ramas y las características utilizadas en cada división [40]. Aunque esta técnica es un poco compleja para el usuario, ya que graficar un número grande (por ejemplo 100) árboles de decisión que pueden ser muy diferentes, no le ayuda mucho a la comprensión del modelo.

En unos casos [41], se incorporó una herramienta destacada en este ámbito, Random Forest Similarity Map (RFMap), la cual ofrece una representación visual interactiva y escalable, diseñada para interpretar tanto el comportamiento global como las decisiones locales de un modelo de RF. RFMap permite a los usuarios analizar visualmente las rutas de decisión utilizadas por el modelo, aprovechando técnicas de reducción de dimensionalidad en 2D para mantener la interpretabilidad sin sacrificar el contexto global del bosque. Sin embargo, la escalabilidad visual sigue siendo un desafío significativo cuando se trata de representar modelos con miles de árboles de decisión; y su enfoque se limita a la representación de reglas y rutas de decisión, lo que podría no capturar todos los aspectos críticos del modelo.

Algunos autores [42] emplearon algoritmos de selección de características para identificar y eliminar variables irrelevantes, lo cual no solo simplifica el modelo, sino que también puede mejorar su rendimiento al eliminar el ruido. Esta técnica es particularmente útil en *data sets* con alta dimensionalidad donde muchas características pueden resultar inútiles. Limitar la profundidad de los árboles en ensambles, como sugieren en otra investigación [43], es una estrategia también posible para simplificar el modelo. La restricción de la profundidad evita que los árboles se vuelvan excesivamente complejos y propensos a sobre ajustarse a los datos de entrenamiento.

Es crucial destacar las innovaciones recientes en XAI aplicadas a modelos de ensamble. Un enfoque emergente es el uso de Explainable Boosting Machines (EBM) [44], que combinó la interpretabilidad de los modelos lineales con la capacidad predictiva de los de ensamble. EBM se utiliza para crear modelos altamente interpretables y precisos, facilitando su adopción en aplicaciones críticas como la salud y las finanzas. Este método ha demostrado ser efectivo en la identificación de patrones complejos y en la explicación de las decisiones del modelo, proporcionando una transparencia que es crucial para la confianza de los usuarios.

Además, se han desarrollado técnicas avanzadas como FairXGBoost, que integra consideraciones de equidad en el proceso de modelado de XGBoost [45]. FairXGBoost ajusta las ponderaciones de las características para mitigar sesgos y asegurar decisiones más justas, asunto particularmente relevante en dominios sensibles como la concesión de créditos y la selección de personal. Estas innovaciones no solo mejoran la interpretabilidad de los modelos de ensamble, sino que también abordan la solución a preocupaciones éticas.

La Tabla 5 compara los hallazgos clave de técnicas XAI aplicadas a modelos de ensamble:

Tabla 5. Técnicas XAI aplicadas a modelos de ensamble

Técnica XAI	Precisión	Interpretabilidad	Eficiencia	Aplicabilidad
SHAP	Alta	Alta	Moderada	Optimizar sistemas críticos
LIME	Media	Alta	Alta	Optimizar sistemas críticos
Árboles de decisión	Alta	Media	Alta	Transparencia de modelos
Valores de permutación	Alta	Baja	Baja	Reducción de sesgos en IA
EBM	Alta	Alta	Alta	Transparencia de modelos
FairXGBoost	Alta	Alta	Moderada	Equidad algorítmica

F. Respuesta a las preguntas de investigación

A continuación, se presentan las respuestas a las preguntas de investigaciones definidas en la Tabla 1, tras realizar el análisis de los artículos seleccionados.

1) ¿Qué innovaciones con enfoque XAI se han usado en modelos de aprendizaje automático basados en ensambles (*bagging*, *boosting*) para problemas de regresión y clasificación y cuáles han sido las ventajas y desventajas que se han observado en su uso? Las innovaciones con enfoque XAI aplicadas a modelos de aprendizaje automático basados en ensambles para problemas de regresión y clasificación, como *bagging* y *boosting*, han buscado mejorar la transparencia e interpretabilidad de estos modelos [46]. Las técnicas más destacadas incluyen la importancia de características basada en permutaciones, los diagramas de dependencia parcial (PDP) y los gráficos de importancia SHAP. Estas permiten identificar qué características son más relevantes para el modelo y cómo contribuyen a las decisiones de clasificación o regresión [47]. La importancia de características basadas en permutaciones permite evaluar el impacto de cada una al medir la variación en el rendimiento del modelo cuando se permutan los valores de una característica específica. Los PDP proporcionan una representación visual de la relación entre una o dos características y la predicción promedio del modelo, ayudando a identificar patrones y tendencias. Los gráficos de importancia SHAP ofrecen explicaciones consistentes y aditivas sobre la contribución de cada característica a las predicciones individuales, facilitando la comprensión de las interacciones complejas entre características.

Una desventaja significativa es que las explicaciones globales proporcionadas por estas técnicas pueden ser insuficientemente detalladas para capturar la influencia específica de las características en casos particulares, especialmente en presencia de relaciones no lineales o interacciones complejas entre características, lo que puede limitar la utilidad de estas explicaciones en aplicaciones que requieren una interpretación detallada a nivel local [48].

Entre las desventajas de las explicaciones locales se destaca que estos métodos pueden ser computacionalmente costosos, sobre todo cuando se aplican a modelos de ensamble con un gran número de estimadores. Además, pueden ser difíciles de interpretar cuando se utilizan ensambles con modelos de base complejos, como árboles de decisión profundos.

Sumado a esto, se han propuesto técnicas de visualización para mostrar la estructura y el comportamiento de los ensambles como la representación gráfica de árboles de decisión o la visualización de la influencia de características en la clasificación. Dentro de las ventajas de estos métodos se destaca que las visualizaciones pueden ayudar a comprender la estructura y la contribución de cada modelo de base en el ensamble facilitando la interpretación y el análisis; y en lo relativo a las desventajas, se resalta que las visualizaciones pueden ser limitadas en la cantidad de información que pueden mostrar.

2) ¿Qué trabajos previos transformaron un bosque de árboles **de un ensamble *bagging* o *boosting* en un solo árbol de decisión?** En el contexto de la transformación de un bosque de árboles de un ensamble *bagging* o *boosting* en un solo árbol de decisión es importante tener en cuenta que esta no es una transformación comúnmente realizada, debido a las diferencias fundamentales entre los ensambles y los árboles de decisión individuales. Los ensambles —como los bosques aleatorios (Random Forest) o los métodos de *boosting* como XGBoost— están diseñados para combinar múltiples árboles de decisión y obtener resultados más precisos y robustos. Sin embargo, trabajos previos han explorado la posibilidad de transformar ensambles en un solo árbol de decisión con el objetivo de simplificar y aumentar la interpretabilidad del modelo resultante.

En uno de estos, los autores propusieron un enfoque para aproximar el modelo XGBoost, un ensamble basado en *gradient boosting*, a un árbol de decisión interpretable. El objetivo era conseguir un modelo simplificado y comprensible que mantuviera un rendimiento cercano al del ensamble original. Este método implicó tres etapas clave: 1) la poda del conjunto para reducir el número de hojas, 2) la extracción de reglas conjuntivas del conjunto podado y 3) la organización de estas reglas en una estructura de árbol de decisión jerárquico [12].

El objetivo del algoritmo presentado consistió en utilizar el modelo entrenado para generar un nuevo árbol que aproximara el rendimiento predictivo del modelo original. Este método se basó en la premisa de que tanto los árboles de decisión como los bosques de decisiones podían representarse como un conjunto finito de reglas conjuntivas, extendiendo así un trabajo anterior, adaptándolo a Gradient Boosting Decision Trees (GBDT) y permitiendo ahora a los usuarios controlar mejor el equilibrio entre el rendimiento predictivo y la complejidad del modelo.

En el otro, se presentó un método para transformar un bosque de decisiones, que es un ensamble basado en *bagging*, en un único árbol de decisión interpretable. El enfoque buscó preservar la estructura de decisión del bosque original mientras proporcionaba una representación más comprensible con el fin de facilitar la interpretación y comprensión del modelo, preservando al mismo tiempo su capacidad predictiva [37].

Esta propuesta partió de un conjunto de datos con n ejemplos, m características y c clases diferentes, un bosque de decisiones agrega K funciones aditivas y mapea un vector de características m-dimensional a un vector de probabilidades c-dimensional. La propuesta del método consistió en construir un nuevo

árbol que aproximara el comportamiento del bosque original, organizando las reglas en una estructura que permitiera predicciones rápidas y comprensibles para instancias no vistas.

La premisa central de este método era que tanto los bosques de decisiones como los árboles de decisión podían ser representados como conjuntos disjuntos de reglas conjuntivas. Es importante señalar que no consideró dependencias entre diferentes árboles por lo que es más adecuado para bosques de decisiones independientes como Random Forests o Rotation Forests.

Estos dos trabajos fueron relevantes en el estado del arte de XAI, ya que abordaron el desafío de convertir modelos de ensamble en árboles de decisión interpretables desde el enfoque más claro de sustitución. Sin embargo, hay pocos estudios disponibles en este campo, lo que confirma la importancia de investigaciones adicionales para mejorar la interpretabilidad de los ensambles *bagging* y *boosting*.

4. CONCLUSIONES Y TRABAJO FUTURO

Los estudios seleccionados permiten destacar que la explicabilidad en la inteligencia artificial (XAI) es un campo de investigación en rápido crecimiento que busca mejorar la comprensión de los modelos de aprendizaje automático y sus decisiones. En particular, en el contexto de modelos de ensamble como *bagging* y *boosting* se han realizado esfuerzos para aplicar técnicas de XAI, mejorar su interpretabilidad y aumentar la confianza en estos al facilitar la validación, la explicación de los resultados y la detección de sesgos y errores. Sin embargo, también existen limitaciones y desafíos en el uso de XAI en modelos de ensamble como complejidad computacional, interpretación inconsistente e irrelevancia de las explicaciones.

Así pues, algunas direcciones para trabajos futuros se relacionan con la necesidad de optimizar los métodos de transformación, la creación de métricas para evaluar la explicabilidad, la integración de técnicas de visualización interactivas y la exploración de técnicas híbridas que combinen enfoques de interpretación global y local.

CONTRIBUCIÓN DE LOS AUTORES

Fabian-Camilo Martinez-Silva: Metodología; Investigación; Análisis formal; Redacción-borrador original; Redacción-revisión y edición

Carlos-Alberto Cobos-Lozada: Conceptualización; Análisis formal; Supervisión; Obtención de financiación; Redacción-revisión y edición.

AGRADECIMIENTOS

Agradecemos al Grupo de Investigación y Desarrollo en Tecnologías de la Información (GTI) de la Universidad del Cauca. A la Universidad del Cauca por financiar parcialmente el desarrollo de esta investigación

REFERENCIAS

[1] D. Minh, H. X. Wang, Y. F. Li, T. N. Nguyen, "Explainable artificial intelligence: a comprehensive review," *Artificial Intelligence Review*, vol. 55, no. 5, pp. 3503-3568, 2022. https://doi.org/10.1007/s10462-021-10088-y

[2] C. Moreira, Y. L. Chou, M. Velmurugan, C. Ouyang, R. Sindhgatta, P. Bruza, "LINDA-BN: An interpretable probabilistic approach for demystifying black-box predictive models," *Decision Support Systems*, vol. 150, e113561, 2021. https://doi.org/10.1016/j.dss.2021.113561

[3] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206-215, May 2019. https://doi.org/10.1038/s42256-019-0048-x

[4] B. A. Kitchenham, D. Budgen, O. P. Brereton, "Using mapping studies as the basis for further research–a participant-observer case study," *Information and Software Technology*, vol. 53, no. 6, pp. 638-651, Jun. 2011. https://doi.org/10.1016/j.infsof.2010.12.011

[5] J. Hatwell, M. M. Gaber, R. Muhammad Atif Azad, "Gbt-hips: Explaining the classifications of gradient boosted tree ensembles," *Applied Sciences (Switzerland)*, vol. 11, no. 6, e2511, Mar. 2021. https://doi.org/10.3390/app11062511

[6] S. Ali *et al*., "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence," *Information Fusion*, vol. January, e101805, Nov. 2023. https://doi.org/10.1016/j.inffus.2023.101805

[7] A. V. Konstantinov, L. V. Utkin, "Interpretable machine learning with an ensemble of gradient boosting machines," *Knowledge-based Systems*, vol. 222, e106993, Jun. 2021. https://doi.org/10.1016/j.knosys.2021.106993

[8] A. Ghose, B. Ravindran, "Interpretability with accurate small models," *Frontiers in Artificial Intelligence*, vol. 3, e3, Feb. 2020. https://doi.org/10.3389/frai.2020.00003

[9] E. S. Ortigossa, T. Gonçalves, L. G. Nonato, "EXplainable Artificial Intelligence (XAI)–From Theory to Methods and Applications," *IEEE Access*, vol. 12, e15, Jun. 2024.

[10] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, P. M. Atkinson, "Explainable artificial intelligence: an analytical review," *WIREs Data Mining and Knowledge Discovery*, vol. 11, no. 5, pp. 1-13, Jul. 2021. https://doi.org/10.1002/widm.1424

[11] F. A. Khalifa, H. M. Abdelkader, A. H. Elsaid, "An analysis of ensemble pruning methods under the explanation of Random Forest," *Information System*, vol. 120, e102310, Feb. 2024. https://doi.org/10.1016/j.is.2023.102310

[12] O. Sagi, L. Rokach, "Approximating XGBoost with an interpretable decision tree," *Information Sciences (NY)*, vol. 572, pp. 522-542, Sep. 2021. https://doi.org/10.1016/j.ins.2021.05.055

[13] Y. Zhou, G. Hooker, "Interpreting models via single tree approximation," *arXiv preprint*, Oct. 2016. https://doi.org/10.48550/arXiv.1610.09036

[14] A. Van Assche, H. Blockeel, "Seeing the forest through the trees: Learning a comprehensible model from an ensemble," in *18th European Conference on Machine Learning*, Warsaw, Poland, 2007, pp. 418-429.

[15] C. Yang, A. Rangarajan, S. Ranka, "Global model interpretation via recursive partitioning," in *IEEE 20th International Conference on High Performance Computing and Communications*, IEEE, Exeter, Reino Unido, 2018, pp. 1563-1570.

[16] M. Nagahisarchoghaei *et al*., "An Empirical Survey on Explainable AI Technologies: Recent Trends, Use-Cases, and Categories from Technical and Application Perspectives," *Electronics (Switzerland)*, vol. 12, no. 5, e1092, Feb. 2023. https://doi.org/10.3390/electronics12051092

[17] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, W. Samek, "Explainable AI Methods-A Brief Overview," in *XX AI - Beyond Explainable AI*. Cham: Springer International Publishing, 2022, pp. 13-38. https://doi.org/10.1007/978-3-031-04083-2_2

[18] M. T. Ribeiro, S. Singh, C. Guestrin "Why should i trust you? Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, San Diego, EE.UU., 2016, pp. 1135-1144.

[19] S. M. Lundberg, S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, e35, 2017.

[20] Á. Delgado-Panadero, B. Hernández-Lorca, M. T. García-Ordás, J. A. Benítez-Andrades, "Implementing local-explainability in Gradient Boosting Trees: Feature Contribution," *Information Sciences (NY)*, vol. 589, pp. 199-212, Apr. 2022. https://doi.org/10.1016/j.ins.2021.12.111

[21] C. Molnar, *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*, 2020.

[22] F. Doshi-Velez, B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint*, 2017.

[23] D. Alvarez-Melis, T. S. Jaakkola, "On the robustness of interpretability methods," *arXiv preprint*, 2018.

[24] M. P. Neto, F. V. Paulovich, "Explainable matrix - Visualization for global and local interpretability of random forest classification ensembles," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 1427-1437, Feb. 2021. https://doi.org/10.1109/TVCG.2020.3030354

[25] D. Mazumdar, M. P. Neto, F. V. Paulovich, "Random Forest similarity maps: A scalable visual representation for global and local interpretation," *Electronics (Switzerland)*, vol. 10, no. 22, e2862, Nov. 2021. https://doi.org/10.3390/electronics10222862

[26] H. Löfström, T. Löfström, U. Johansson, C. Sönströd, "Investigating the impact of calibration on the quality of explanations," *Annals of Mathematics and Artificial Intelligence*, vol. 23, e10472, Mar. 2023. https://doi.org/10.1007/s10472-023-09837-2

[27] A. Sudjianto, J. Qiu, M. Li, J. Chen, "Linear iterative feature embedding: an ensemble framework for an interpretable model," *Neural Computing and Applications*, vol. 35, no. 13, pp. 9657-9685, Mar. 2023. https://doi.org/10.1007/s00521-023-08204-w

[28] A. Altmann, L. Toloşi, O. Sander, T. Lengauer, "Permutation importance: a corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340-1347, May 2010. https://doi.org/10.1093/bioinformatics/btq134

[29] L. Antwarg, C. Galed, N. Shimoni, L. Rokach, B. Shapira, "Shapley-based feature augmentation," *Information Fusion*, vol. 96, pp. 92-102, Aug. 2023. https://doi.org/10.1016/j.inffus.2023.03.010

[30] M. Louhichi, R. Nesmaoui, M. Marwan, M. Lazaar, "Shapley Values for Explaining the Black Box Nature of Machine Learning Model Clustering," *Procedia Computer Science*, vol. 220, pp. 806-811, 2023. https://doi.org/10.1016/j.procs.2023.03.107

[31] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189-1232, Oct. 2001. https://doi.org/10.1214/aos/1013203451

[32] A. V Konstantinov, and L. V Utkin, "Interpretable machine learning with an ensemble of gradient boosting machines," *Knowledge-based Systems*, vol. 222, e106993, Jun. 2021. https://doi.org/10.1016/j.knosys.2021.106993

[33] M. P. Neto, F. V Paulovich, "Explainable matrix-visualization for global and local interpretability of random forest classification ensembles," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 1427-1437, Feb. 2020. https://doi.org/10.1109/TVCG.2020.3030354

[34] A. Adadi, M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138-52160, 2018. https://doi.org/10.1109/ACCESS.2018.2870052

[35] G. Yang, Q. Ye, J. Xia, "Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond," *Information Fusion*, vol. 77, pp. 29-52, Jan. 2022. https://doi.org/10.1016/j.inffus.2021.07.016

[36] A. Rai, "Explainable AI: from black box to glass box," *Journal of the Academy of Marketing Science*, vol. 48, no. 1, pp. 137-141, 2020. https://doi.org/10.1007/s11747-019-00710-5

[37] O. Sagi, L. Rokach, "Explainable decision forest: Transforming a decision forest into an interpretable tree," *Information Fusion*, vol. 61, pp. 124-138, Sep. 2020. https://doi.org/10.1016/j.inffus.2020.03.013

[38] V. Hassija *et al.*, "Interpreting black-box models: a review on explainable artificial intelligence," *Cognitive Computation*, vol. 16, no. 1, pp. 45-74, 2024. https://doi.org/10.1007/s12559-023-10179-8

[39] T. Speith, "A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods," in *ACM International Conference Proceeding Series*, pp. 2239-2250, Nueva York, EE.UU., 2022. https://doi.org/10.1145/3531146.3534639

[40] K. Dedja, F. K. Nakano, K. Pliakos, C. Vens, "BELLATREX: Building Explanations through a LocaLly AccuraTe Rule EXtractor," *IEEE Access*, vol. 11, pp. 41348-41367, 2023. https://doi.org/10.1109/ACCESS.2023.3268866

[41] D. Mazumdar, M. P. Neto, F. V Paulovich, "Random Forest similarity maps: a scalable visual representation for global and local interpretation," *Electronics (Basel)*, vol. 10, no. 22, e2862, Nov. 2021. https://doi.org/10.3390/electronics10222862

[42] I. Guyon, A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157-1182, Mar. 2003.

[43] T. Hastie, R. Tibshirani, J. H. Friedman, J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Nueva York, EE.UU.: Springer, 2009.

[44] H. Nori, S. Jenkins, P. Koch, R. Caruana, "Interpretml: A unified framework for machine learning interpretability," *arXiv preprint*, 2019.

[45] S. Ravichandran, D. Khurana, B. Venkatesh, N. U. Edakunni, "Fairxgboost: Fairness-aware classification in xgboost," *arXiv preprint*, 2020.

[46] B. Zhang, J. Zhu, H. Su, "Toward the third-generation artificial intelligence," *Science China Information Sciences*, vol. 66, no. 2, pp. 1-19, Jan. 2023. https://doi.org/10.1007/s11432-021-3449-x

[47] R. Marcinkevičs, J. E. Vogt, "Interpretable and explainable machine learning: A methods-centric overview with concrete examples," *WIREs Data Mining and Knowledge Discovery*, vol. 13, no. 3, pp. 1-32, Feb. 2023. https://doi.org/10.1002/widm.1493

[48] L. A. Cox, "Information structures for causally explainable decisions," *Entropy*, vol. 23, no. 5, e601, May 2021. https://doi.org/10.3390/e23050601

Fabian-Camilo Martínez-Silva, Carlos-Alberto Cobos-Lozada

REVISIÓN SISTEMÁTICA DE LAS HERRAMIENTAS DE INTELIGENCIA ARTIFICIAL EXPLICABLE USADAS EN MÉTODOS DE ENSAMBLE

Explainable Artificial Intelligence tools used in ensemble methods: A systematic literature review

Revista Facultad de Ingeniería
vol. 33, núm. 70, e18078, 2024
Universidad Pedagógica y Tecnológica de Colombia,
ISSN: 0121-1129
ISSN-E: 2357-5328

DOI: https://doi.org/10.19053/01211129.v33.n70.2024.18078