# *Descifrando el enigma: Interpretabilidad de modelos de caja negra*

REPOSITORIO: https://github.com/dandiasua02/Interpretabilidad

Alejandro García Sánchez-Hermosilla
*dpto. Ciencias de la Computación e Inteligencia Artificial*
*Universidad de Sevilla*
Sevilla, España
alegarsan11@alum.us.es

Daniel Diáñez Suárez
*dpto. Ciencias de la Computación e Inteligencia Artificial*
*Universidad de Sevilla*
Sevilla, España
dandiasua@alum.us.es

**El objetivo principal de este trabajo se ha basado en desarrollar y aplicar un modelo XAI de LIME para explicar cómo funcionan los diferentes modelos de caja negra que hemos desarrollado y entrenado previamente. Dichos modelos son: redes neuronales y RandomForest.**

**Como conclusión de este proyecto podemos destacar la importancia de los modelos XAI, ya que puede ser fundamental para entender las tareas que realizan los modelos de caja negra; es decir, modelos de los que no se conoce su funcionamiento.**

## I. Introducción

En la época actual de la inteligencia artificial y el aprendizaje automático, los modelos de caja negra han demostrado un rendimiento excepcional en una amplia gama de aplicaciones, desde diagnósticos médicos hasta sistemas de recomendación personalizados. Sin embargo, debido al avance de la tecnología, han surgido problemas en la comprensión de los modelos de caja negra debido a qué no se hace explícito su comportamiento; es decir, no sabemos cómo funcionan por dentro.

Dado este problema, surge las XAI (Inteligencia Artificial Explicable) que aborda dichos problemas mediante el uso de técnicas de aprendizaje automático. En este proyecto nos centraremos en un modelo que utiliza técnicas aplicadas a posteriori (Post-hoc), el modelo en concreto que hemos desarrollado es el modelo LIME (Local Interpretable Model-agnostic Explanations).

Dicho prototipo ofrece unas explicaciones sobre un modelo. Estas explicaciones se obtienen mediante un modelo subrogado (RIDGE) y que se entrena con una predicción individual del modelo a explicar, cuyos datos de la muestra a predecir son perturbados aleatoriamente dentro del intervalo de valores de ese atributo.

El objetivo perseguido en este trabajo es dar explicaciones a modelos cuyo comportamiento es opaco con el uso de un modelo LIME. Pero no ha sido trabajo fácil ya que hemos encontrado problemas durante el desarrollo de este proyecto, algunos de estos problemas son los siguientes. Entrenamiento de modelos RandomForest, ya que es un modelo nuevo para nosotros, tuvimos que buscar información e informarnos para poder entrenar dicho modelo. Reformulación de datos para las distintas tareas, para poder realizar clasificación multiclase. La diferenciación de modelos para la obtención de explicaciones, debido a problemas con los resultados obtenidos en el modelo LIME con los distintos modelos.

En este documento se explicará la metodología seguida para la realización de este trabajo y se estructurará de la siguiente manera: **Preliminares**, **Metodología**, **Resultados** y **Conclusiones**.

## II. Preliminares

En esta sección se hace una breve introducción de los métodos que hemos empleado en este trabajo.

A. Redes neuronales: son modelos computacionales inspirados en el funcionamiento del cerebro humano. En nuestro trabajo hemos utilizado las redes neuronales para realizar tareas de modelado y predicción. Estas redes están compuestas por capas de neuronas interconectadas, donde cada capa procesa y transforma la información antes de pasarla a la siguiente capa. Es importante destacar que las redes neuronales requieren de un conjunto de datos de entrenamiento para aprender los patrones y relaciones en los datos y ajustar los pesos de las conexiones entre las neuronas. [1]

B. RandomForest: es un algoritmo de aprendizaje automático basado en el concepto de "bosques aleatorios", consiste en un conjunto de árboles de decisión individuales que trabajan en paralelo para realizar predicciones. En el contexto de nuestro

trabajo, hemos utilizado RandomForest como un algoritmo de clasificación. [2]

C. Técnicas de clasificación: En nuestro trabajo, hemos explorado diferentes técnicas de clasificación, incluyendo clasificación binaria y clasificación multiclase. Además, hemos aplicado técnicas específicas de interpretación adaptadas a cada tipo de clasificación para mejorar la comprensión y explicación de los modelos. [3]
   - La clasificación binaria se refiere a la tarea de clasificar instancias en dos clases o categorías distintas.
   - La clasificación multiclase implica la clasificación en más de dos clases.

## III. METODOLOGÍA

Esta sección se dedica a la descripción del método implementado en el trabajo.

El método que hemos implementado en nuestro código es el algoritmo LIME y consiste en generar un conjunto de permutaciones aleatorias de la muestra original x, permutando selectivamente un número k de atributos en cada perturbación. Para cada permutación generada, se calcula una representación numérica que capture la diferencia entre la perturbación y la muestra original. Además, se mide la distancia entre la muestra original y la perturbación generada.

Posteriormente, se obtienen las predicciones Y' utilizando el modelo f sobre las muestras perturbadas. A continuación, se entrena un modelo de regresión ridge utilizando las representaciones generadas para predecir las predicciones Y'. Cada muestra perturbada se pondera con su correspondiente distancia medida anteriormente.

El objetivo de este método es obtener un modelo ridge que permita comprender cómo los cambios en los atributos afectan a las predicciones del modelo f. Al ponderar las muestras perturbadas en función de sus distancias, se busca asignar mayor peso a aquellas perturbaciones que son más similares a la muestra original.

**Algorithm LIME: pseudocódigo**

N es el n´umero de permutaciones a realizar
f es el modelo a explicar
X′ ← {} muestras perturbadas
R ← {} representaciones
W ← {} las distancias entre la muestra x y sus perturbaciones
**for** 1 to N **do**
    Selecciona k atributos aleatoriamente
        x ′ ← una perturbaci´on de x donde se perturban los k atributos anteriores.
    w ← la distancia entre x y x ′
    r ← la representaci´on de x ′ respecto a x
    X′ ← X′ ∪ x ′
    R ← R ∪ r
    W ← R ∪ w
**end for**
Y ′ ← f(X′ ) las predicciones de las perturbaciones
G ← modelo ridge entrenado con R para predecir Y ′ y ponderando cada muestra con W
**return** los parámetros de G

**Algorithm LIME: formula matemática**

$$\xi(x) = \underset{g \in G}{argmin}\, \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

## IV. Resultados

Para los diferentes modelos hemos tenido en cuenta diferentes formas de trabajar para tener homogeneidad en los resultados, para ello hemos modificado el modelo LIME a nuestras necesidades para que no interfiriese los resultados de las predicciones en el funcionamiento del modelo XAI.

Los resultados obtenidos según las explicaciones del modelo son los coeficientes del modelo subrogado que indican la importancia de cada atributo a la hora de realizar una tarea de predicción con una muestra. En nuestro caso, para medir los resultados, hemos hecho uso de una serie de métricas (identidad, separabilidad, estabilidad, selectividad, coherencia, completitud y congruencia). Cada una de estas métricas nos indican datos relevantes sobre las explicaciones del modelo, a continuación, explicaremos cada una de estas métricas:

| **Métrica** | **Descripción** |
|---|---|
| Identidad | Muestras idénticas deben tener explicaciones idénticas |
| Separabilidad | Muestras que no son idénticas no pueden tener explicaciones idénticas |
| Estabilidad | Los objetos cercanos deben tener explicaciones similares |
| Selectividad | Cuanto mayor sea la variación entre las características importantes, menos preciso será la predicción. |
| Coherencia | En un modelo coherente, si tomamos decisions iguales, obtendremos explicaciones similares. |
| Completitud | Evalúa el porcentaje de error de explicación con respecto al error de predicción. |
| Congruencia | Ayuda a capturer la variabilidad de la coherencia |

## V. Conclusiones

Los experimentos realizados han demostrado que el método propuesto es capaz de proporcionar una explicación razonablemente precisa del modelo. Los coeficientes del modelo ridge obtenido han permitido comprender cómo los cambios en los atributos afectan a las predicciones del modelo, lo cual es crucial para la interpretación y validación del mismo.

En cuanto a las conclusiones, se ha observado que el método de perturbación aleatoria es una estrategia efectiva para explorar y analizar diversos modelos. Proporciona información valiosa sobre la importancia de los atributos en las predicciones y puede ayudar a identificar posibles sesgos o relaciones no lineales en el modelo.

Por último, el método de perturbación aleatoria ha demostrado ser una herramienta prometedora para la explicación de modelos y la interpretación de decisiones de diversos modelos. Facilita una base sólida para comprender y validar los modelos y puede ser utilizado en diversos campos de aplicación para mejorar la confianza y transparencia de los sistemas.

## Referencias


[1] https://stackoverflow.com/search?q=neural+networks

[2] https://stats.stackexchange.com/search?q=randomforest

[3] Página web del curso IA de Ingeniería del Software. https://www.cs.us.es/docencia/aulavirtual/course/view.php?id=30