PyData

Explicabilidad algorítmica – SHAP

Paloma Megías Mesa

28 de abril de 2022

instituto de ingeniería del conocimiento

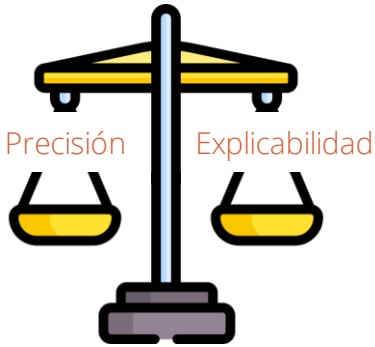
www.iic.uam.es

EXPLICABILIDAD

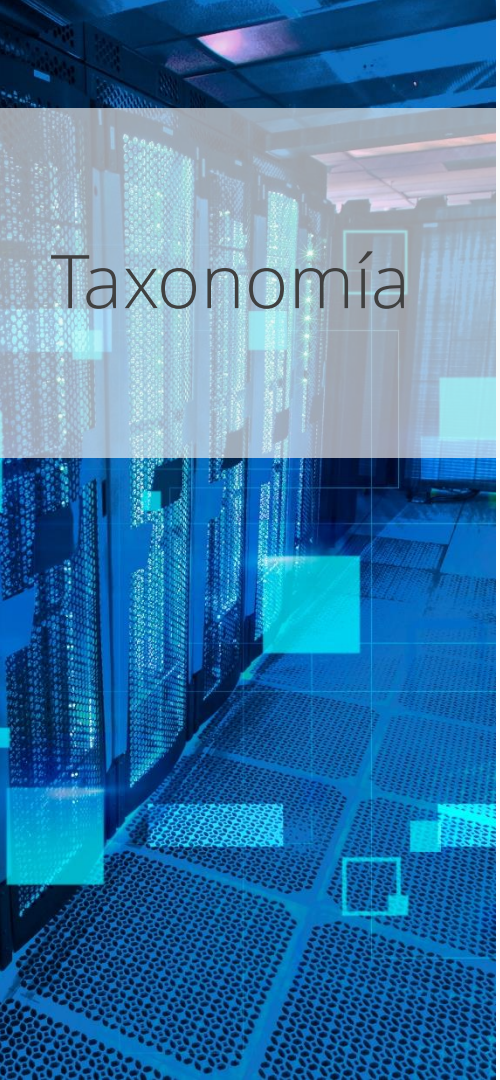
Es el grado en que un humano puede comprender la causa de una decisión.

Machine Learning

¿Qué está haciendo mi modelo?



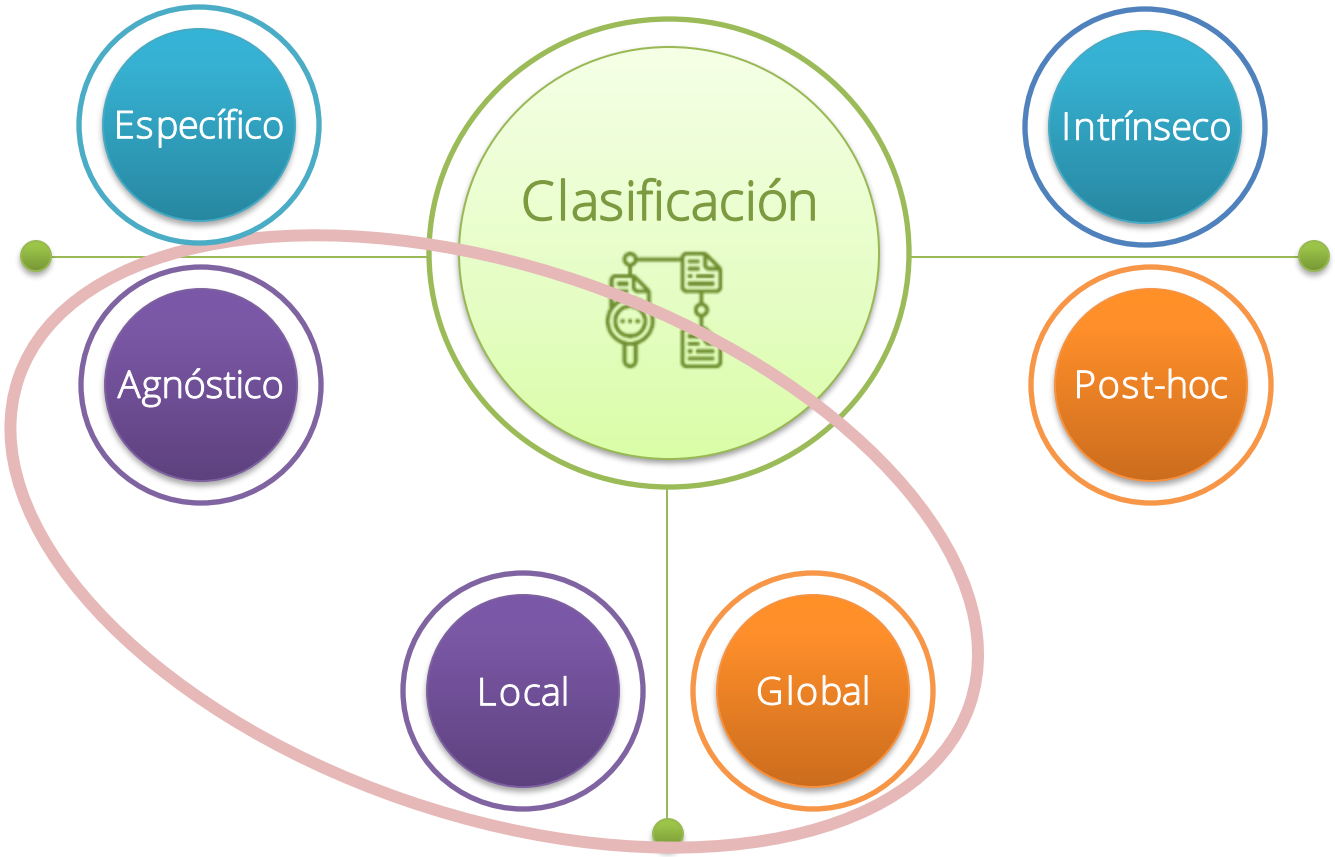
Miller, Tim. *"Explanation in artificial intelligence: Insights from the social sciences." arXiv Preprint arXiv:1706.07269. (2017).*



Taxonomía

C. Molnar. (2018). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Accessed: Jun. 6, 2018.

Taxonomía



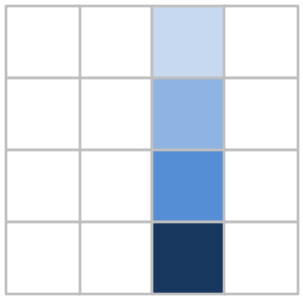
C. Molnar. (2018). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Accessed: Jun. 6, 2018.

Métodos globales

Permutation importance

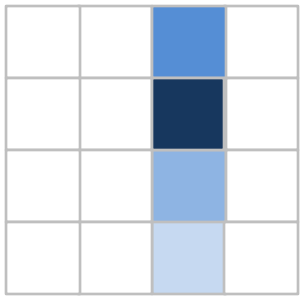
Importancia de una variable = Aumento en el error de predicción al permutar la variable.

Original



variable V

Permutado



variable V

Error ⟷ Error

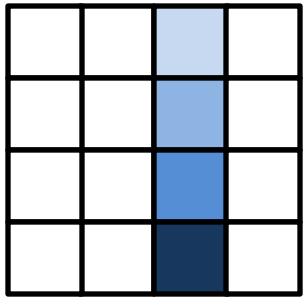
Problema:
Asume que las variables no están relacionadas entre sí.

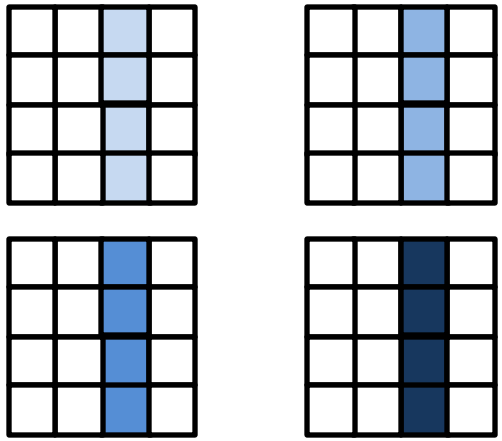
Métodos globales

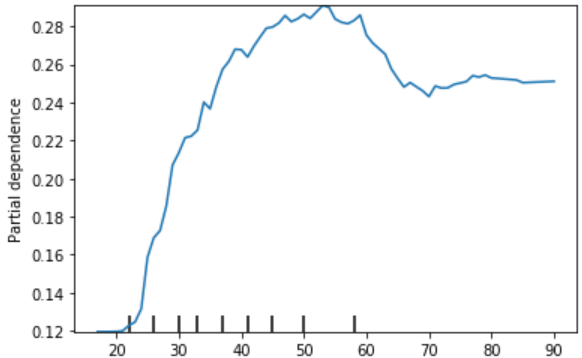
Gráficas de dependencias parciales

Impacto de una variable sobre las predicciones del modelo

Original







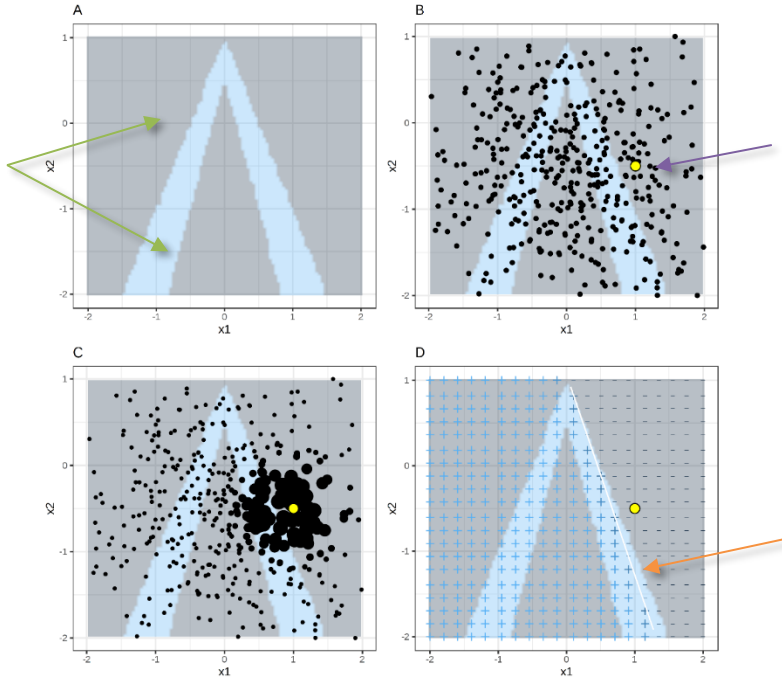
Métodos locales

LIME

Entrena modelos sustitutos locales para explicar las predicciones individuales.

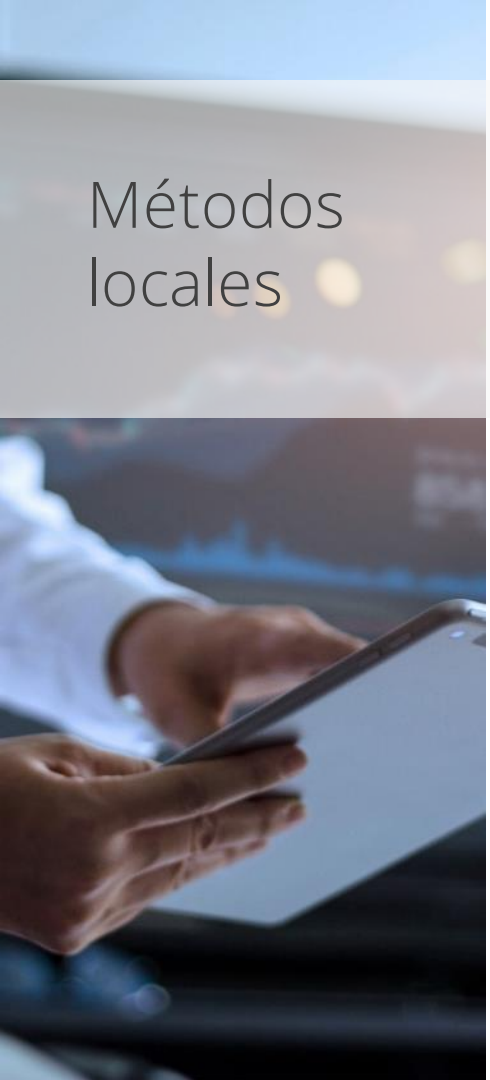
- Se asigna importancia en función en función de la distancia al punto amarillo.
- Ajustamos un modelo lineal para el dato de interés.
- Problema: **Es muy inestable.**

Decisiones del modelo de caja negra.



Dato de interés

Frontera de decisión del modelo lineal



Métodos locales

Valores de Shapley

Es un método para asignar valor a los jugadores en función de su contribución al resultado total. Los jugadores cooperan en una coalición y reciben una cierta ganancia de esta cooperación.

¿Qué jugador es más importante?

Importancia de un jugador = cuánto aporta al resultado cuando se une.

https://en.wikipedia.org/wiki/Shapley_value

Valores de Shapley

Propiedades

Eficiencia: La suma de los valores de Shapley de todos los jugadores es igual al valor total del juego.

Simetría: Si dos jugadores son equivalentes, su valor de Shapley es el mismo.

Aditividad: Si el beneficio del juego se puede descomponer linealmente, los valores de Shapley también.

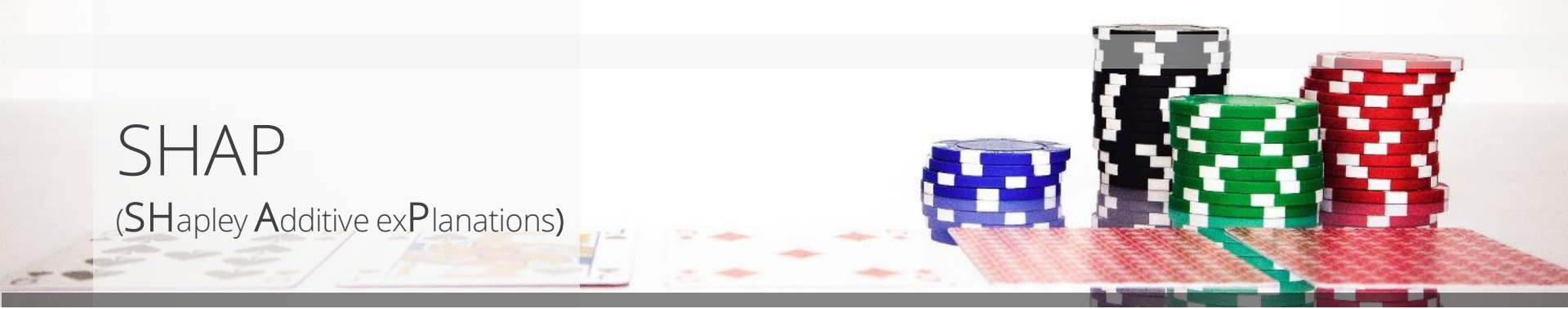
Jugador nulo: Si un jugador nunca aporta nada, su valor de Shapley es 0.

Los valores de Shapley son la única manera de asignar valor a los jugadores de forma que se cumplan estas 4 propiedades simultáneamente.

¿Cómo lo medimos en modelos predictivos?

SHAP

(SHapley Additive exPlanations)

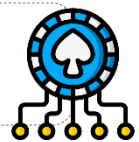


SHAP

Es un método para calcular coeficientes de Shapley en un modelo de explicación local.

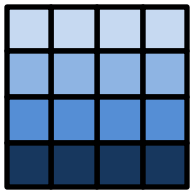
Cada jugador es una variable y el beneficio que nos da el modelo es su rendimiento.

"Cuánto aporta cada variable al resultado del modelo"

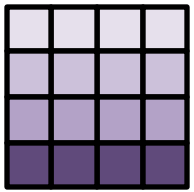


Librería SHAP

Original data



SHAP values



Paloma Megías

Data Scientist

https://www.linkedin.com/in/paloma-m-39075b176

Puedes consultar los artículos de innovación en nuestro Blog:
www.iic.uam.es/blog/

Elementos gráficos de apoyo obtenidos en:
designed by freepik.com
pixabay

C/ Francisco Tomás y Valiente, nº 11
EPS, Edificio B, 5ª planta
UAM Cantoblanco. 28049 Madrid
Tel.: (+34) 91 497 2323