TOULOUSE
LAUTREC

MACHINE LEARNING

Frase para el curso

"All models are wrong,
but some are useful"

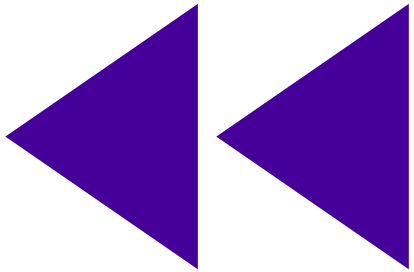
George Edward Pelham Box,
estadístico británico.
(1919 – 2013)

Objetivo de la clase

Aplicar algoritmos de regresión: El estudiante aprenderá a implementar varios tipos de algoritmos de regresión, incluyendo la regresión lineal simple, múltiple y no lineal.

Palabras clave: aprendizaje supervisado, regresión lineal, regresión múltiple, regresión no lineal, evaluar modelo

Recordando conceptos ...



Recordando algunos conceptos ...

1. X: variable de entrada, variable predictora, características, features, variable independiente, observaciones.

2. Y: variable respuesta, outcome, variable dependiente, variable objetivo, respuesta, objetivo, target, clase.

3. Para entrenar un modelo de Machine Learning se debe tener datos de entrenamiento, pruebas y validación.

   Una posible distribución entre los dataset es:
   70 – 80% entrenamiento
   20 – 30% pruebas
   0 – 10% de validación

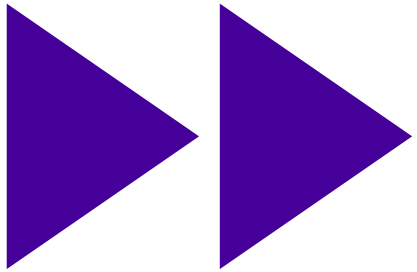
Recordando algunos conceptos ...

4. Aprendizaje supervisado es cuando se tienen bien definidos los predictores y la respuesta.

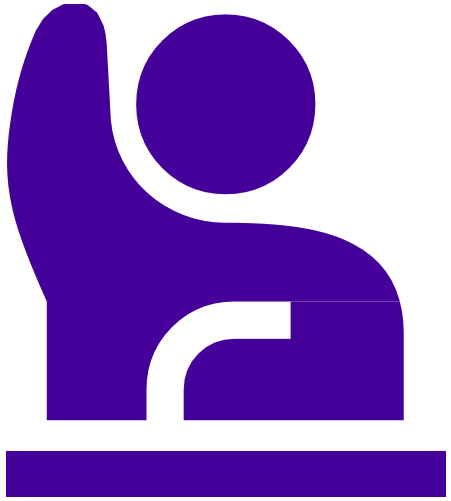
5. La mayoría de los problemas de aprendizaje supervisado pueden formularse formalmente en términos de predicción de una respuesta, pero la predicción por sí sola no suele ser el objetivo principal del análisis.

6. Por ejemplo, muchas aplicaciones de la regresión lineal en las ciencias tienen como **objetivo principal comprender cómo las entradas de un sistema impulsan las salidas**; una "caja negra" extremadamente complicada que ofreciera predicciones puras no sería muy útil en sí misma.

Continuamos ...



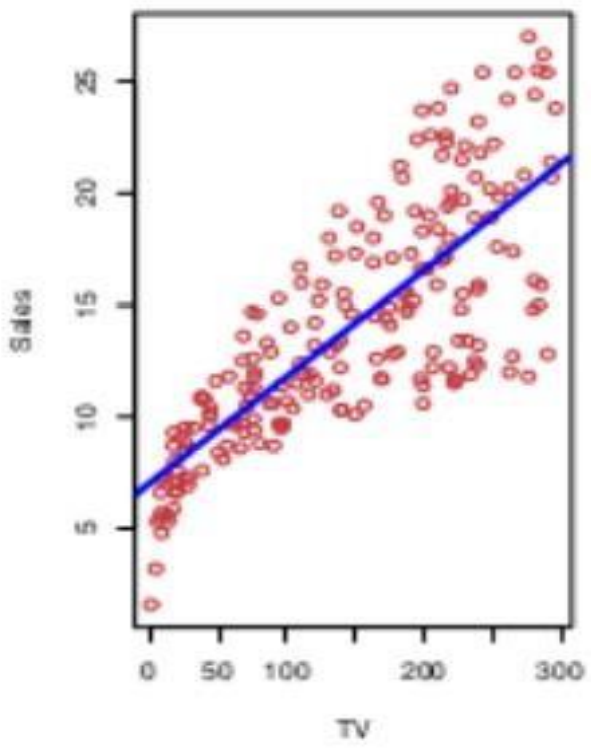
Para participar:

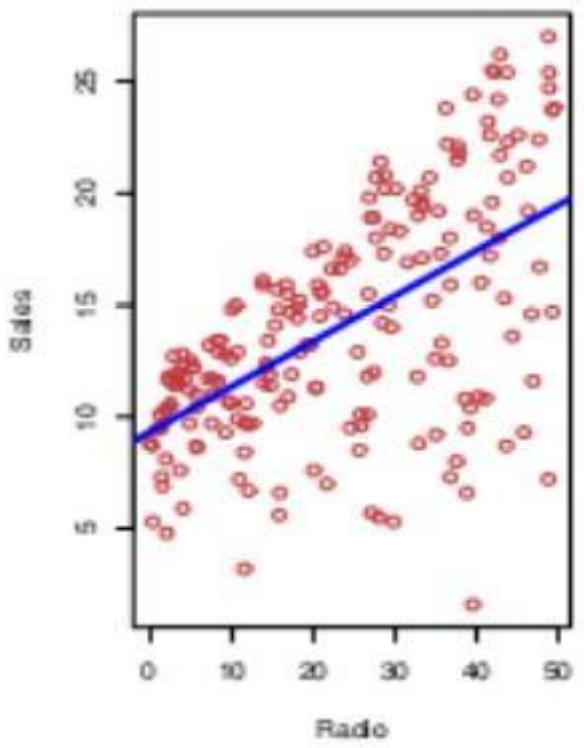


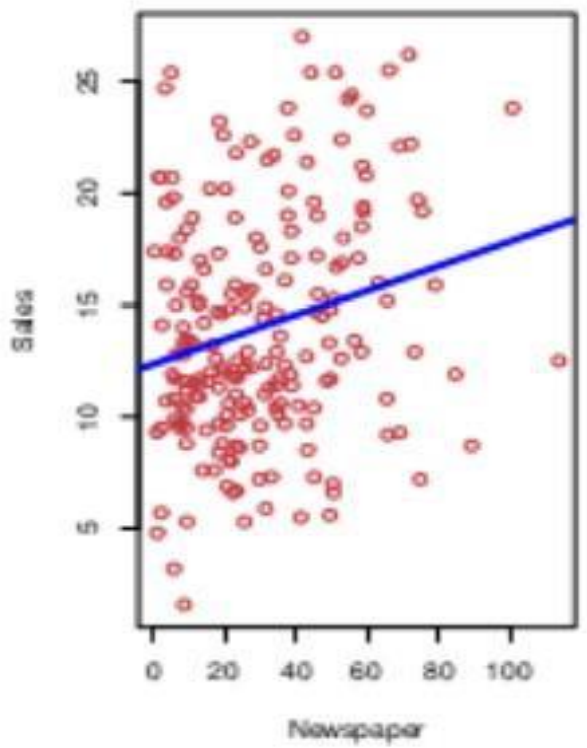
22487105

https://www.menti.com/aly65xuoifqy

Regresión lineal para estimar las ventas







$$\text{Sales} \approx f(\text{TV}, \text{Radio}, \text{Newspaper})$$

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \qquad Y = f(X) + \epsilon$$

Regresión lineal

1. ¿Existe alguna relación entre el presupuesto de publicidad y las ventas?
2. ¿Qué tan fuerte es la relación entre el presupuesto de publicidad y las ventas?
3. ¿Cuál medio de comunicación contribuye más a las ventas?
4. ¿Con qué tanto acierto podemos predecir las ventas a futuro?
5. ¿Existe una relación lineal?
6. ¿Existe alguna sinergia entre los medios de comunicación?

Regresión lineal

Supongamos que nuestra tabla tiene $n+1$ columnas (todas numéricas): X_1, X_2,...,X_n y Y. Es decir, hay $n+1$ características, y queremos explicar la variable Y *linealmente* a través de las otras variables. Se propone que la explicación se puede ver como

$$\hat{y} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

Lo que queremos es hallar $\alpha, \beta_1, \beta_2, \cdots, \beta_n$ que hagan que el error, ϵ, sea mínimo.

$$Y = f(X, \beta) + \varepsilon$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i$$

- Y es la variable dependiente.
- X son las variables independientes.
- f es la función que describe la relación entre Y y X.
- β son los parámetros del modelo que se ajustan durante el proceso de entrenamiento.
- ε representa la variabilidad no explicada por el modelo.

Predictores cualitativos

$$x = \begin{cases} 1 & \text{Si la i-ésima persona es mujer} \\ 0 & \text{Si la i-ésima persona es hombre} \end{cases}$$

El modelo resultante es:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{Si la i-ésima persona es mujer} \\ \beta_0 + \varepsilon_i & \text{Si la i-ésima persona es hombre} \end{cases}$$

One Hot Encoding

Permite separar las variables predictoras categóricas en diferentes variables predictoras.

Crea una columna binaria para cada categoría y devuelve una matriz dispersa. De esta manera el modelo solamente evaluará dos valores por cada columna.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i + \varepsilon_i$$

$$y = \begin{cases} \beta_0 + \beta_1 + \varepsilon & \text{Si la i-ésima persona es mujer} \\ \beta_0 + \beta_2 + \varepsilon_i & \text{Si la i-ésima persona es hombre} \end{cases}$$

¿Como medir el error de mi modelo?

Error Absoluto Medio

El Error Absoluto Medio (Mean Absolute Error o MAE) se define como:

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Error Cuadrático Medio

El Error cuadrático Medio (Mean Squared Error o MSE)

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

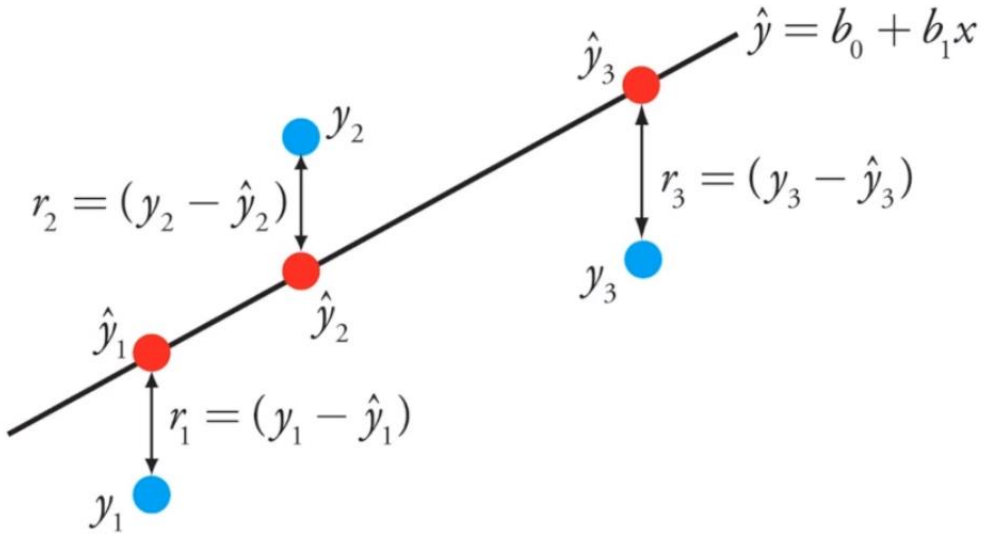
Dado que el MSE se define en unidades al cuadrado, lo cual no es intuitivo (¿dolares cuadrados?), generalmente se usa su raíz.

Raíz del Error Cuadrático Medio

La Raíz del Error Cuadrático Medio (Root Mean Squared Error o RMSE) se diferencia del MSE en que el resultado se puede medir en las mismas unidades que la variable objetivo

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Sin embargo, tiene un problema y es que da más importancia a los errores grandes.



Coeficiente de Determinación

Hay varias formas de definir R2 , pero una de las más sencillas es simplemente la correlacion (definida como la Correlación de Pearson) entre la variable objetivo y las predicciones , elevada al cuadrado.

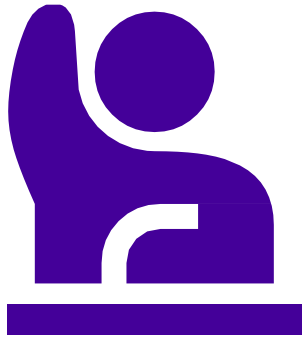
Por eso una medida mejor es el Coeficiente de Determinacion ajustado (Adjusted R-squared), que tiene en consideracion la complejidad del modelo

$$1 - \frac{(1 - R^2)(n - 1)}{(n - k - 1)}$$

Para participar:

¿Cuál de los siguientes modelos es mejor?

Modelo 1	Modelo 2
R^2 = 0.99	R^2 = 0.99
MSE es 1.34	MSE es 1000.89

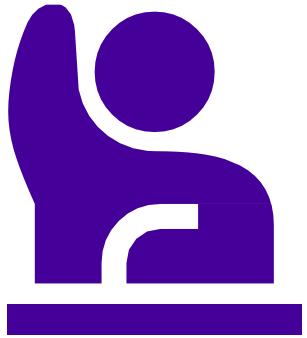


R^2 coeficiente de determinación.
MSE: Error cuadrático medio

Para participar:

¿Cuál de los siguientes modelos es mejor?

Modelo 1	Modelo 2
R^2 = 0.99	R^2 = 0.99
MSE es 1.34	MSE es 1000.89



R^2 coeficiente de determinación.
MSE: Error cuadrático medio

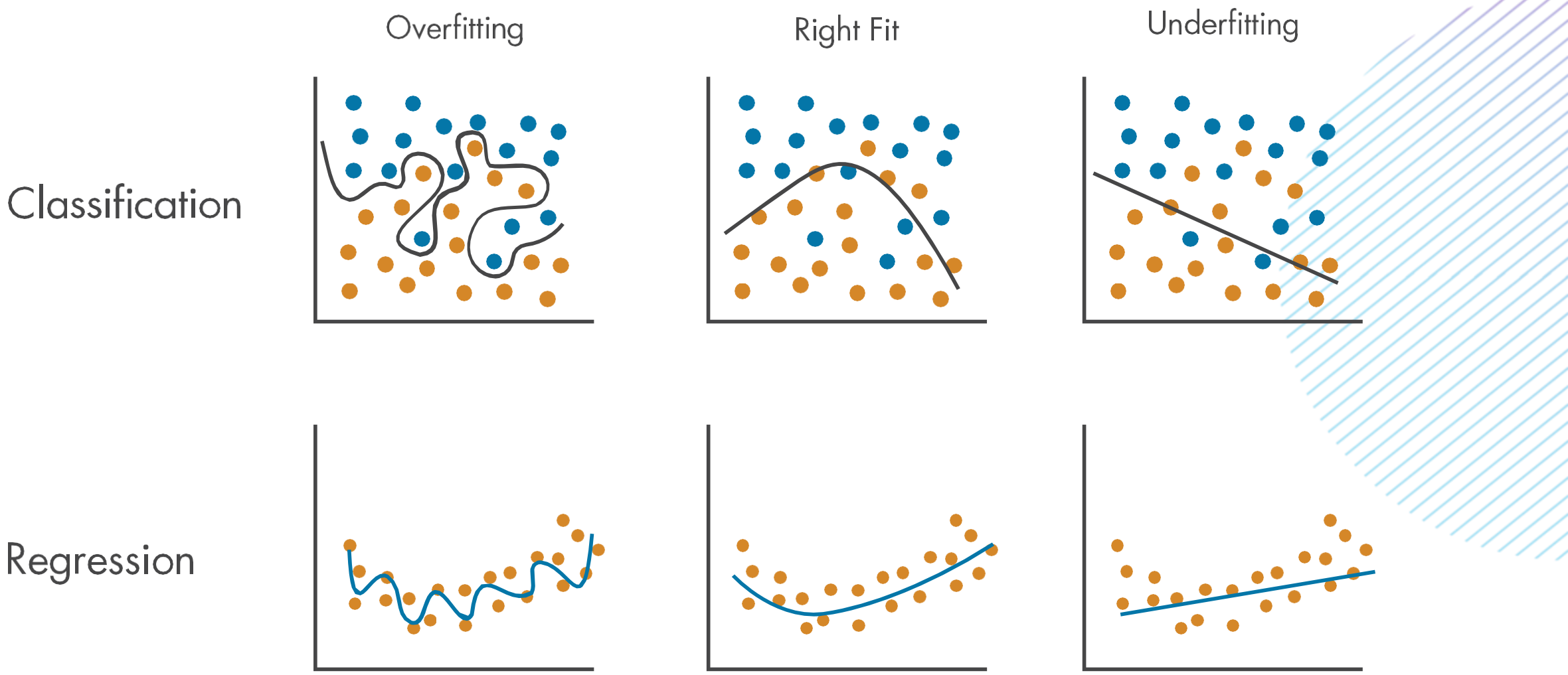
Validación del modelo

- R^2 (coeficiente de determinación) explica cuánta varianza puede explicar el modelo. Variación de los datos.
  Interpretación: Mientras el valor sea más cercano a 1, el modelo explica mejor el conjunto de datos de entrenamiento/test.

- **MSE** (Error cuadrático medio) es más sensible a valores atípicos.
  Por sí solo no permite ver que tan bueno es un modelo.
  Interpretación: Penalizar fuertemente los outliers grandes.

- **MAE** (Error absoluto medio) no es tan sensible a valores atípicos.
  Interpretación: No penalizar demasiado los outliers positivos ni negativos.

Overfitting (sobreajuste) / Underfitting (infrajuste)



Overfitting (sobreajuste) / Underfitting (infrajuste)

Entrenamiento	Test
R^2 = 1.00	R^2 = 0.13
MSE es 0.15	MSE es 4090.67

Si $\boldsymbol{R^2}$ del conjunto de entrenamiento es mucho mayor al de test, entonces hay overfitting.

Overfitting (sobreajuste) / Underfitting (infrajuste)

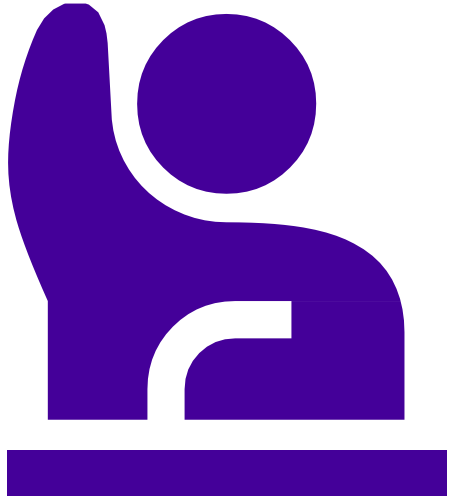
Entrenamiento	Test
$R^2 = 0.17$	$R^2 = 0.13$
MSE es 8000.27	MSE es 4090.67

Si $\boldsymbol{R^2}$ del conjunto de entrenamiento es muy bajo, entonces hay underfitting.

Para participar:

¿Qué es un árbol de decisión?

¿Qué hiperparámetros tiene un árbol de decisión?



XGBoost (eXtreme Gradient Boosting)

"… es una biblioteca optimizada de aumento de gradiente distribuido diseñada para ser altamente eficiente, flexible y portátil. Implementa algoritmos de aprendizaje automático bajo el marco de Gradient Boosting…"

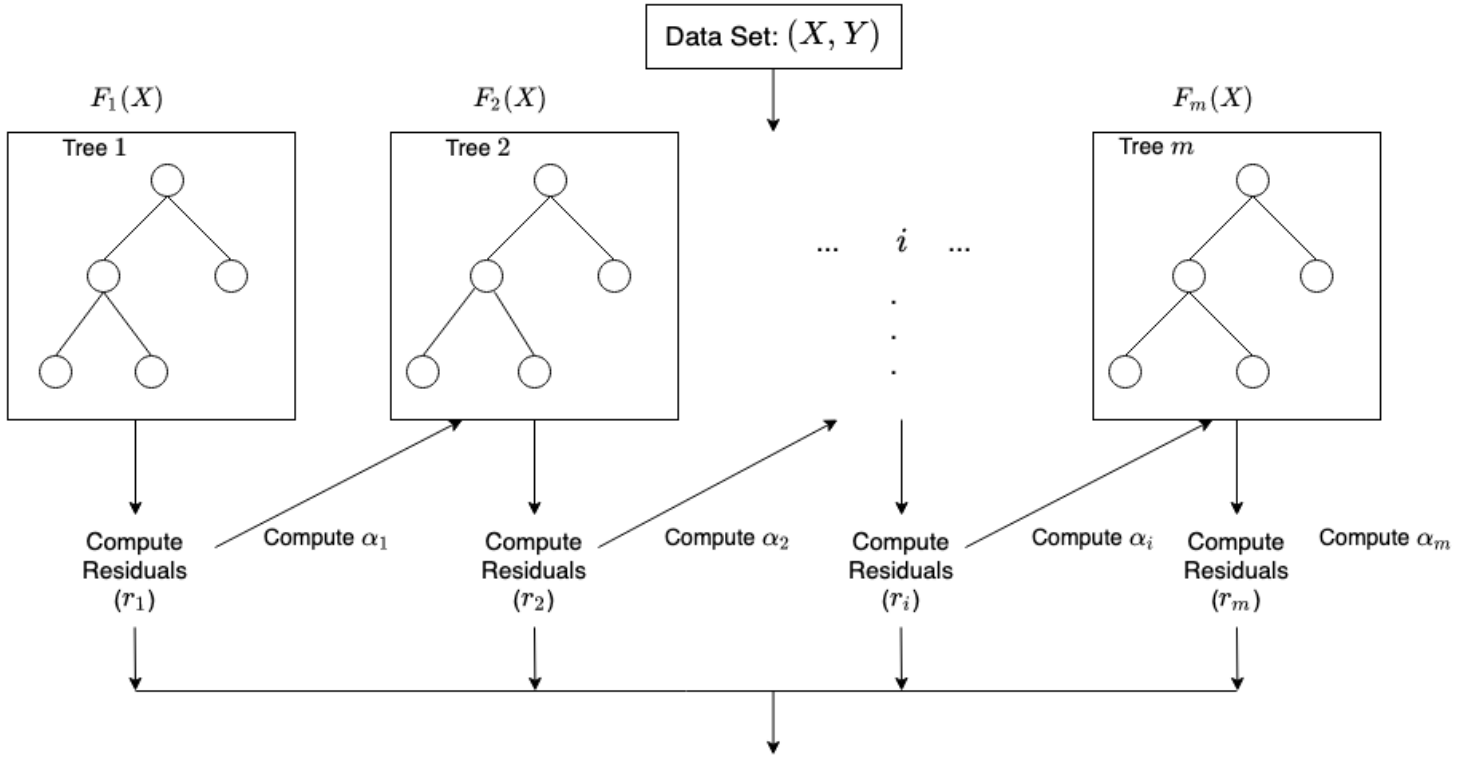
Utiliza un conjunto de árboles de decisión.

A partir del learning rate (taza de aprendizaje) cuánto va a medir los errores para mejorar el modelo.

Es útil para aprendizaje supervisado de regresión y clasificación.

https://xgboost.readthedocs.io/en/stable/

XGBoost (Extreme Gradient Boosting)



$$F_m(X) = F_{m-1}(X) + \alpha_m h_m(X, r_{m-1}),$$

where α_i, and r_i are the regularization parameters and residuals computed with the i^{th} tree respectfully, and h_i is a function that is trained to predict residuals, r_i using X for the i^{th} tree. To compute α_i we use the residuals computed, r_i and compute the following: $arg \min_{\alpha} = \sum_{i=1}^{m} L(Y_i, F_{i-1}(X_i) + \alpha h_i(X_i, r_{i-1}))$ where $L(Y, F(X))$ is a differentiable loss function.

https://docs.aws.amazon.com/es_es/sagemaker/latest/dg/xgboost-HowItWorks.html

XGBoost (eXtreme Gradient Boosting)

objective: Busca minimizar el error durante el entrenamiento.

n_estimators: Número de árboles que llevan a cabo el boosting.

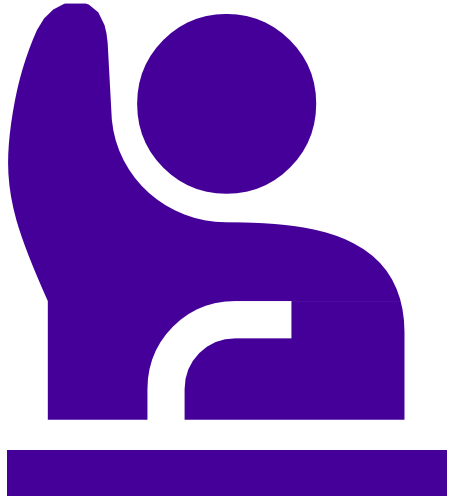
learning_rate: Ratio de aprendizaje. Reduce el tamaño de los pesos del algoritmo.

max_depth: Máxima profundidad que tendrá el árbol. Un valor alto hará más complejo el demolo y posiblemente haya overfitting.

https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/

Para participar:

¿Conoces o has oído de SHAP?



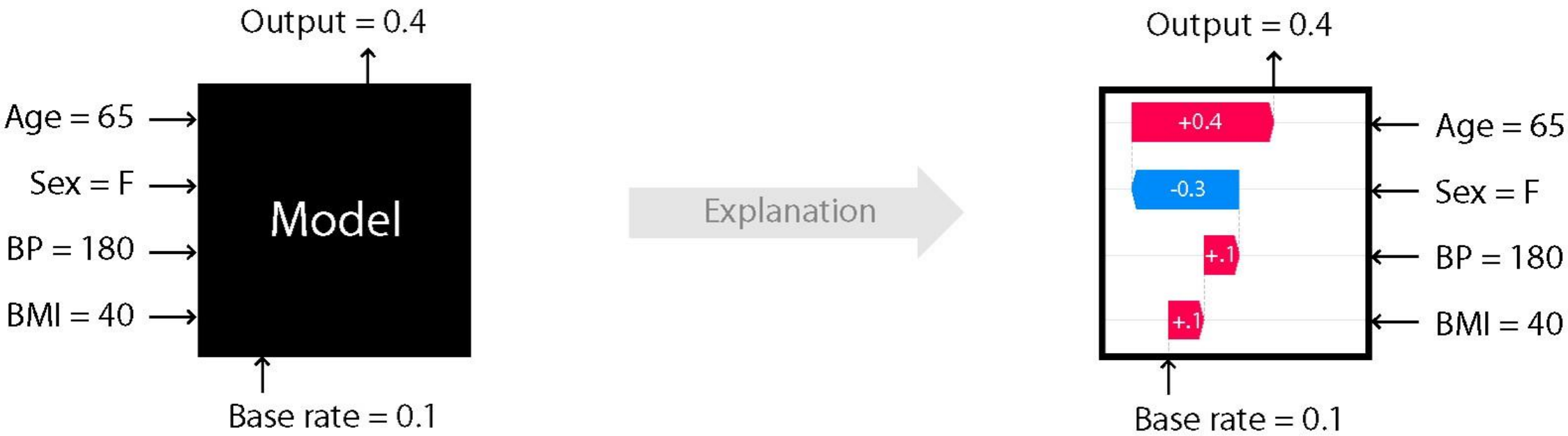
SHAP (SHapley Additive exPlanations)

Permite entender cómo cada característica contribuye a la predicción de la variable objetivo.

Explica las predicciones de modelos de aprendizaje automático.

Facilita la interpretación de modelos complejos.

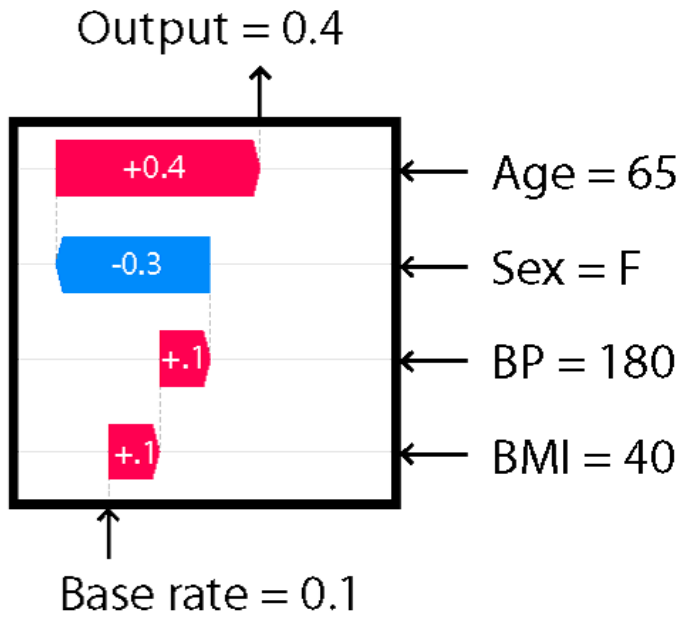
SHAP (SHapley Additive exPlanations)



https://shap.readthedocs.io/en/latest/
https://github.com/shap/shap#tree-ensemble-example-xgboostlightgbmcatboostscikit-learnpyspark-models

SHAP – Interpretación del gráfico

- A mayor valor de Age, el impacto es de manera positiva en el resultado.
- A menor valor de Sex, el impacto es de manera negativamente en el resultado.



https://shap.readthedocs.io/en/latest/
https://github.com/shap/shap#tree-ensemble-example-xgboostlightgbmcatboostscikit-learnpyspark-models

Para participar:

¿Qué interpretas del gráfico?

Alcohol

Acidez volátil

Sulfatos

Dióxido de azufre total

pH

Cloruros

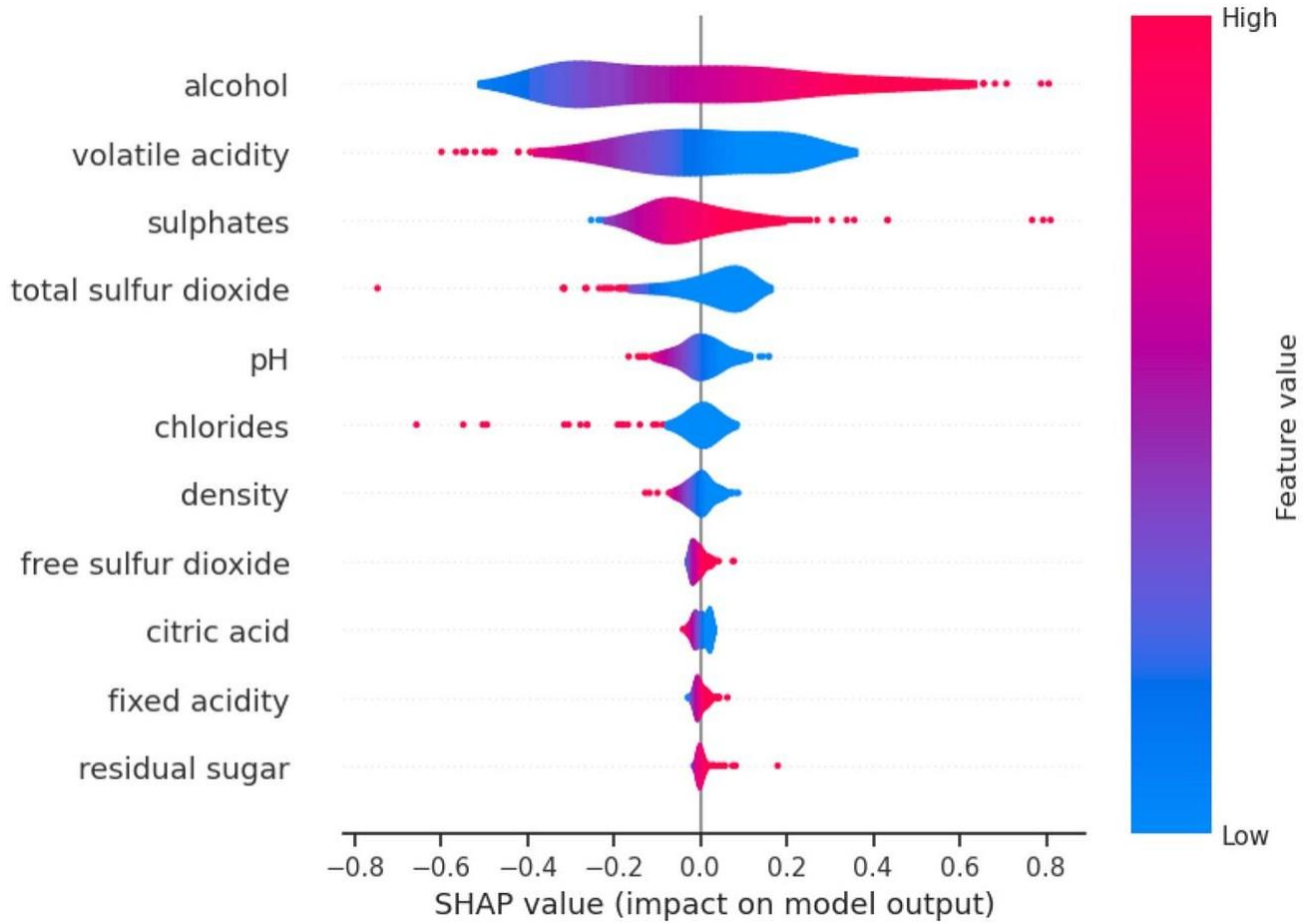
Densidad

Libre de dióxido de azufre

Ácido cítrico

Acidez fija

Azúcar residual



El resultado es la calidad del vino rojo.

https://medium.com/the-power-of-ai/machine-learning-explainability-with-shap-ee60fbfa21e

Para participar:

¿Qué interpretas del gráfico?

Alcohol

Acidez volátil

Sulfatos

Dióxido de azufre total

pH

Cloruros

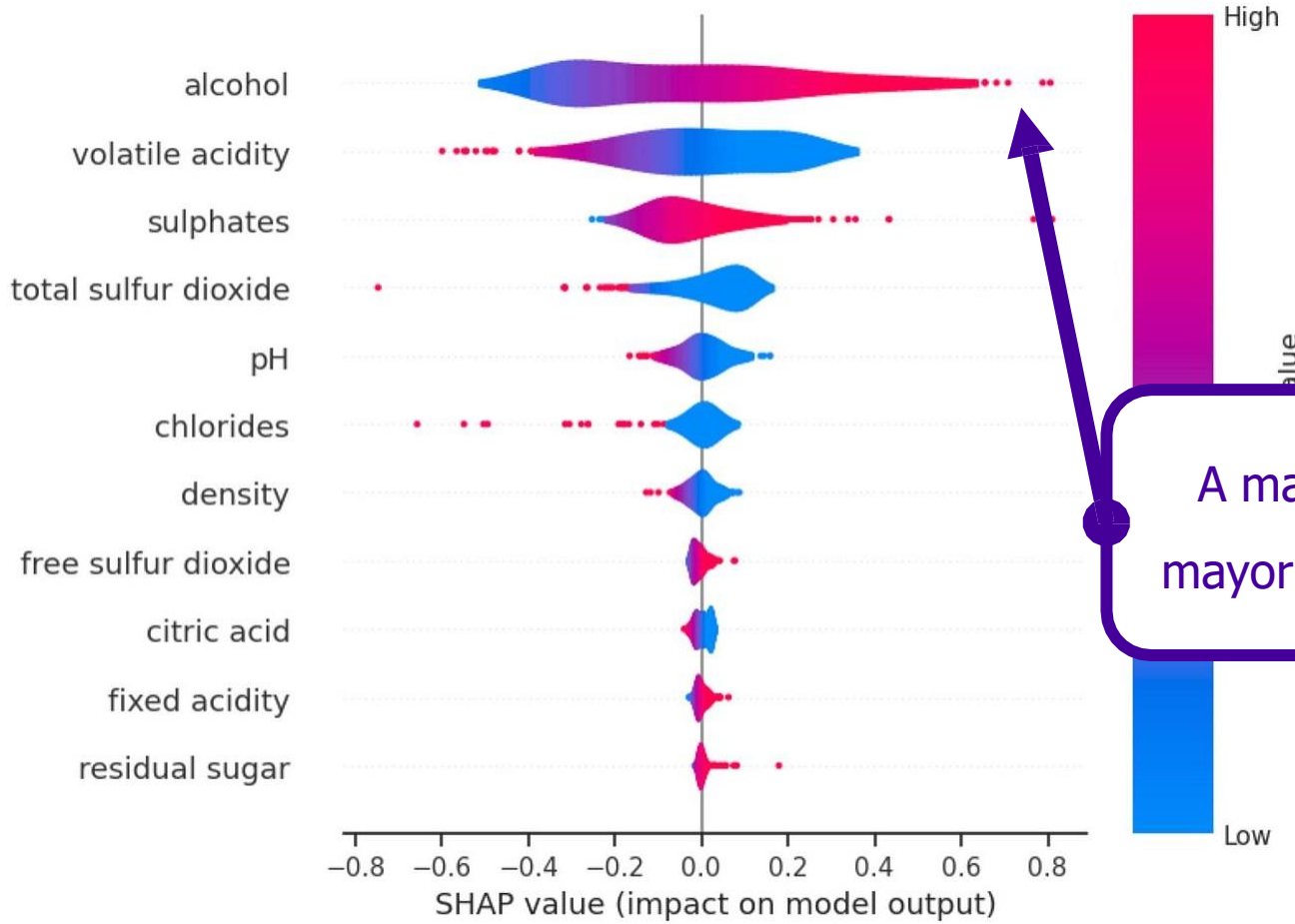
Densidad

Libre de dióxido de azufre

Ácido cítrico

Acidez fija

Azúcar residual



El resultado es la calidad del vino rojo.

A mayor valor de "alcohol", mayor será la calidad del vino.

https://medium.com/the-power-of-ai/machine-learning-explainability-with-shap-ee60fbfa21e

Para participar:

¿Qué interpretas del gráfico?

Alcohol

Acidez volátil

Sulfatos

Dióxido de azufre total

pH

Cloruros

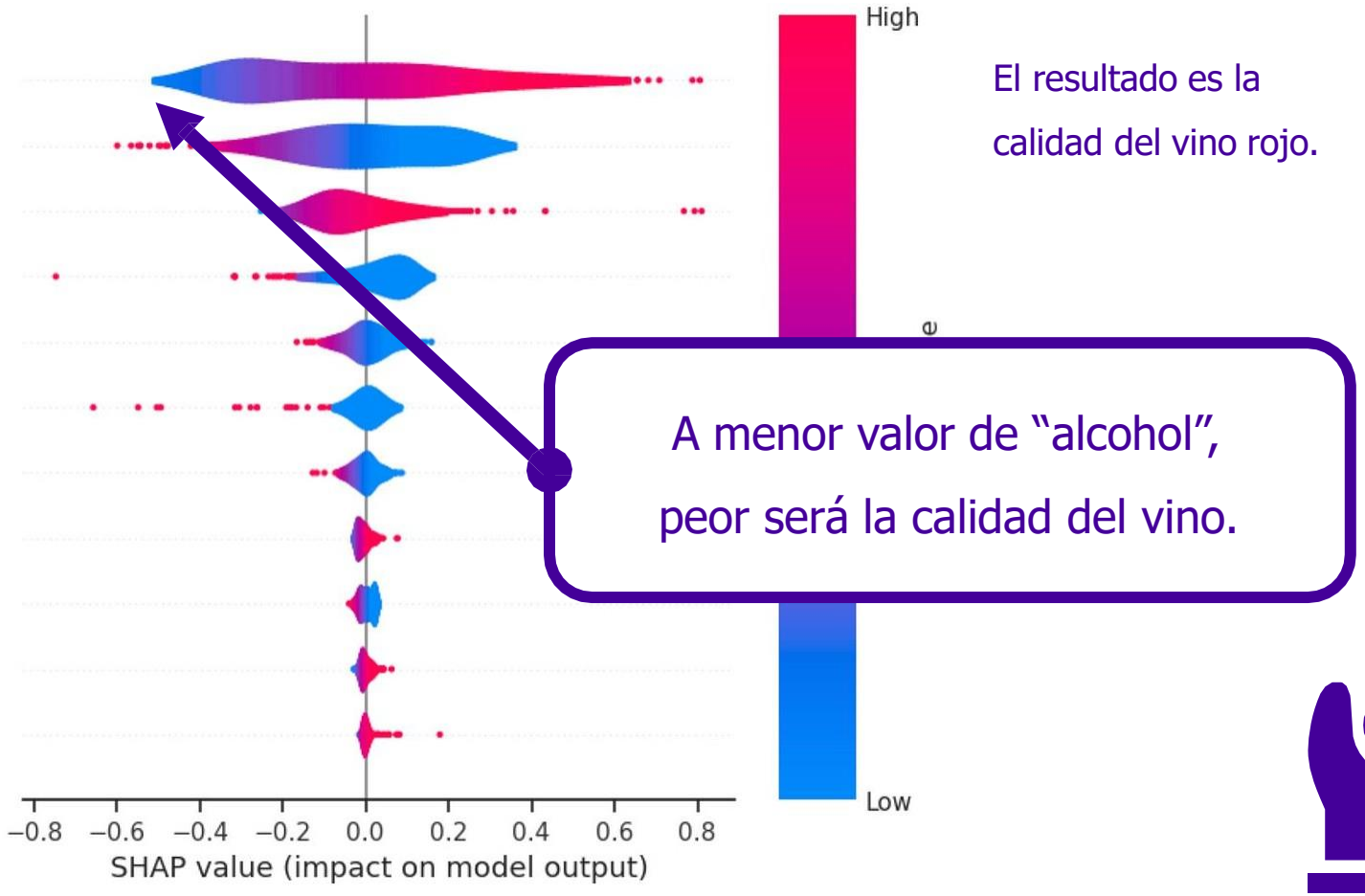
Densidad

Libre de dióxido de azufre

Ácido cítrico

Acidez fija

Azúcar residual



El resultado es la calidad del vino rojo.

A menor valor de "alcohol", peor será la calidad del vino.

https://medium.com/the-power-of-ai/machine-learning-explainability-with-shap-ee60fbfa21e

Para participar:

¿Qué interpretas del gráfico?

Alcohol

Acidez volátil

Sulfatos

Dióxido de azufre total

pH

Cloruros

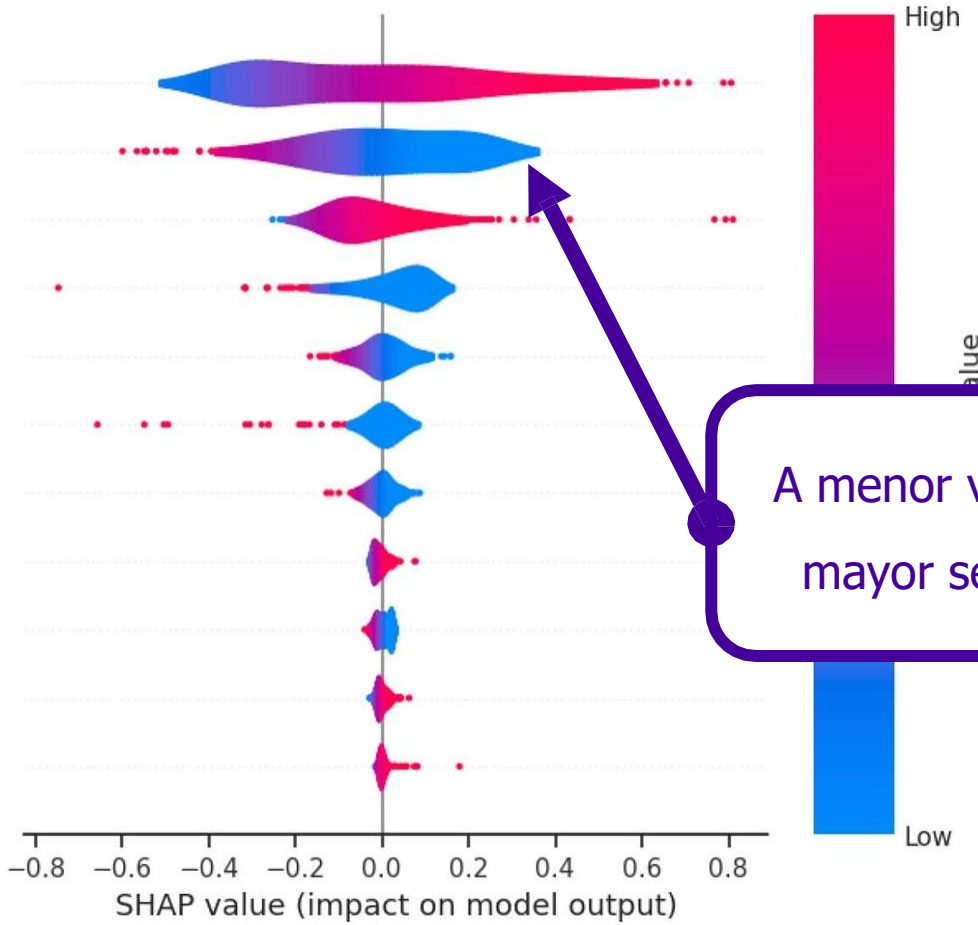
Densidad

Libre de dióxido de azufre

Ácido cítrico

Acidez fija

Azúcar residual



El resultado es la calidad del vino rojo.

A menor valor de "acidez volátil", mayor será la calidad del vino.

https://medium.com/the-power-of-ai/machine-learning-explainability-with-shap-ee60fbfa21e

Para participar:

¿Qué interpretas del gráfico?

Alcohol

Acidez volátil

Sulfatos

Dióxido de azufre total

pH

Cloruros

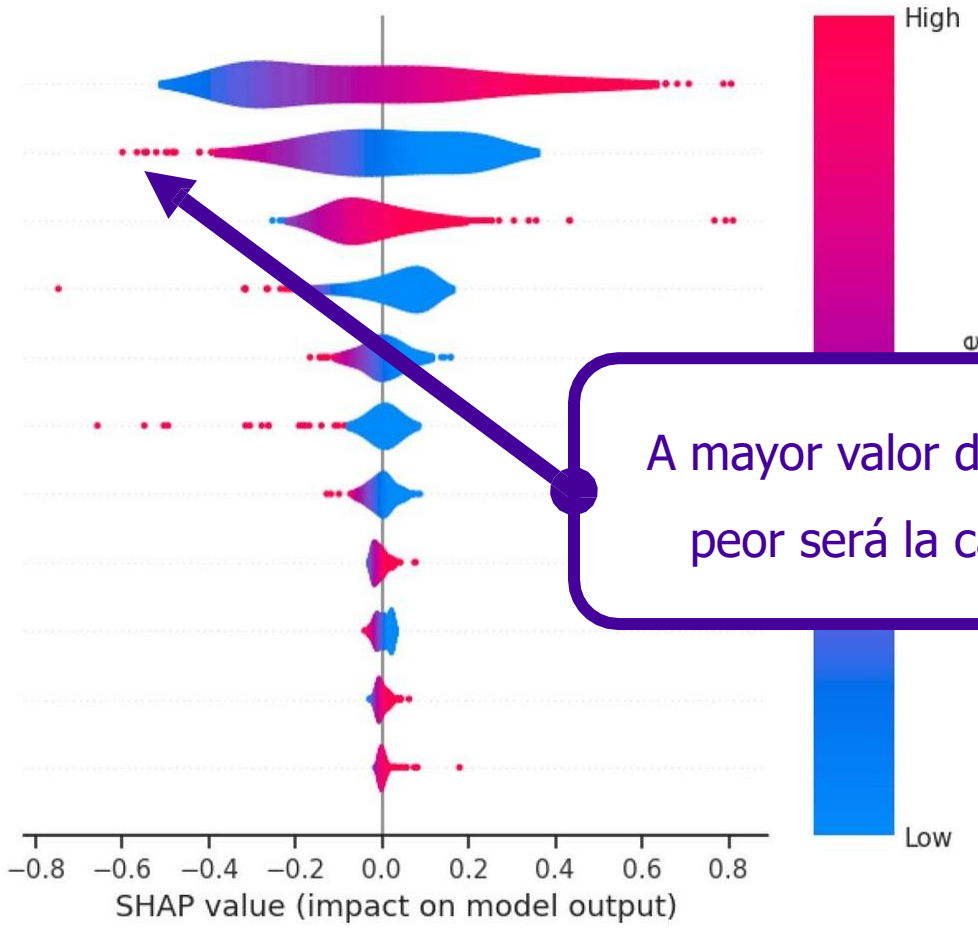
Densidad

Libre de dióxido de azufre

Ácido cítrico

Acidez fija

Azúcar residual

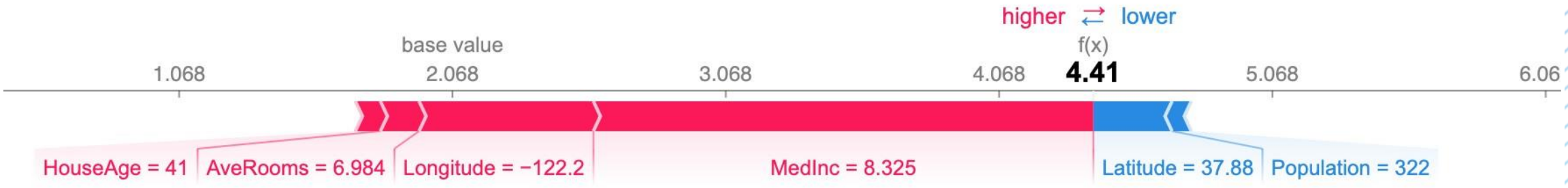


El resultado es la calidad del vino rojo.

A mayor valor de "acidez volátil", peor será la calidad del vino.

https://medium.com/the-power-of-ai/machine-learning-explainability-with-shap-ee60fbfa21e

SHAP – Interpretación de una observación específica



$f(x) = 4.41$

- El valor de MedInc impacta de manera positiva en el resultado.
- El valor de Latitude impacta de manera negativa en el resultado.

https://shap.readthedocs.io/en/latest/

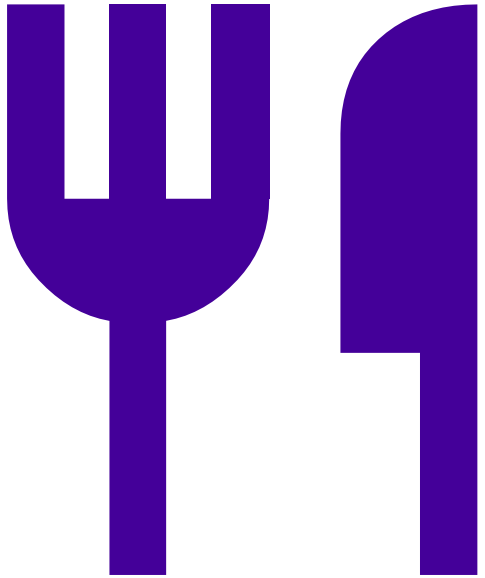
Ejercicio en Google Colab

Predicción de gastos médicos:

- El dataset contiene 1,338 muestras y 18 características.
- Los pacientes tienen entre 18 y 64 años.
- La variable a predecir es **charges**.

https://www.kaggle.com/code/huseyinbaytar/medical-cost-prediction

Pausa de 5 minutos



Ejercicio en Google Colab

Predicción de expectativa de vida.

El dataset contiene 2,938 muestras entre los años 2000 y 2015 de diferentes países.

Predecir la expectativa de vida (variable objetivo: Life Expectancy).

https://www.kaggle.com/code/sadiqshah/bike-store-sales-in-europe/notebook

Taller

1. Resolver los ejercicios de la tarea indicados en el notebook presentado por el profesor.

2. Guardar la respuesta de cada una de las pre
Cargar la tarea en la plataforma antes del 23/01/2024 08:00 p.m. para obtener la nota máxima de 20 puntos.

3. guntas en un archivo con extensión .docx (Word) o PDF o en el Notebook Colab proporcionado por el profesor.

4. **Plazo máximo para la entrega: 27/01/2024 08:00 p.m. La calificación es sobre 15 puntos.**

Evaluación usando Kahoot

1. Ingresar al link de Kahoot que proporcionará el profesor.
2. Hay 10 preguntas. Cada pregunta tiene 4 alternativas. Solamente 1 alternativa es la correcta.
3. Tienen 1 minuto para responder cada pregunta.
4. Esta evaluación se considera como nota de participación.

Cierre de la clase

1. XGBoost permite evaluar diferentes árboles de decisión para obtener el mejor resultado para un modelo de regresión o clasificación.

2. SHAP nos permite identificar cuáles variables tienen mayor impacto en el modelo.

3. One Hot Encoding es un método que permite separar los valores de una variable categórica en diferentes variables predictoras.

4. Completar las encuestas:

   Formulario para conocerte: https://forms.gle/HBnfHHdbrwz1Dtfy7

   ¿Cómo calificas tú la clase?: https://www.menti.com/alz26sgj4g64

Antes de finalizar
¿hay alguna pregunta respecto a la clase?

¿Qué he aprendido el día de hoy en clase?

TOULOUSE
LAUTREC