# Predicción de tarifas de taxi en Nueva York.

## Anticipa tu tarifa.
## Cualquier día, cualquier hora. 🚕


**by Gerardo Jiménez**


Diciembre 2024


1. Recolección y limpieza de datos
2. Feature engineering
3. Prueba de modelos ML
4. Optimización de hiper parámetros
5. SHAP
6. Resultados y conclusiones

# Contenido

# Recolección y Limpieza de Datos

1

## Obtención de datos

Base de datos recuperada de **HuggingFace**, con información de viajes de taxi en Nueva York, entre los años 2009 y 2024.

54M de datos con registros de precio, coordenadas, fecha y hora de los viajes.

2

## Limpieza general

Eliminación de valores nulos y datos incorrectos para garantizar la calidad del conjunto.

3

## Transformación y muestreo de datos

Conversión de formatos y creación de un nuevo archivo CSV con 1M de datos para realizar el análisis.

# Feature engineering

### 🚕 Creación de nuevas variables

Cálculo de datos adicionales como **distancia lineal del viaje** y **ocurrencia de viajes en hora pico**, con base en la información inicial.

Transformación trigonométrica de datos temporales cíclicos para un análisis ML correcto.

### 🚕 Análisis de Componentes Principales (PCA)

Reducción de dimensionalidad a las coordenadas de inicio y fin del viaje, para identificar las variables más influyentes en el precio.

### 🚕 EDA multivariable

Visualización de distribuciones, correlaciones y patrones en los datos de viajes de taxi.

Normalización de variables para asegurar una comparación equitativa entre diferentes escalas.


# Prueba de Modelos de machine learning

### 🚕 Árbol de decisiones (Bagging)

- Error absoluto medio (MAE): 2.20 [USD]
- Ajuste del modelo a la variabilidad de los datos (R2_score): 76.44 %
- Tiempo de ejecución: 7.33 [s]

### 🚕 Random Forest

- Error absoluto medio (MAE): 2.15 [USD]
- Ajuste del modelo a la variabilidad de los datos (R2_score): 77.3 %
- Tiempo de ejecución: 109.42 [s]

### 🚕 AdaBoost

- Error absoluto medio (MAE): 2.06 [USD]
- Ajuste del modelo a la variabilidad de los datos (R2_score): 79.07 %
- Tiempo de ejecución: 219.81 [s]

### 🚕 Gradient Boosting

- Error absoluto medio (MAE): 2.48 [USD]
- Ajuste del modelo a la variabilidad de los datos (R2_score): 64.34 %
- Tiempo de ejecución: 143.80 [s]

### 🚕 XG Boost

- Error absoluto medio (MAE): 2.32 [USD]
- Ajuste del modelo a la variabilidad de los datos (R2_score): 71.46 %
- Tiempo de ejecución: 22.39 [s]

*

Datos de pruebas  con 100,000 datos. Las pruebas fueron realizadas con 1k, 10k, 100k y 1M de datos.

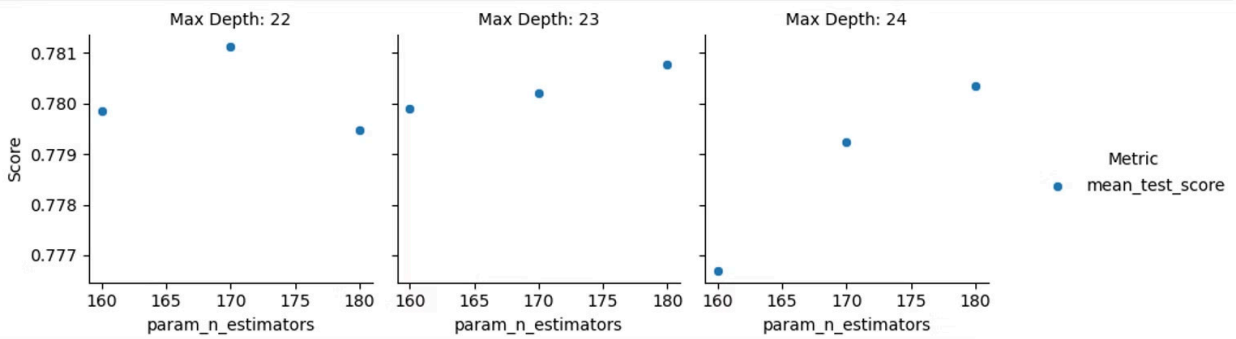
🚕 *Se elige Random Forest por su relación de precisión y tiempo de ejecución.* 🚕

# Optimización de Hiperparámetros

## 🚕 Random Search

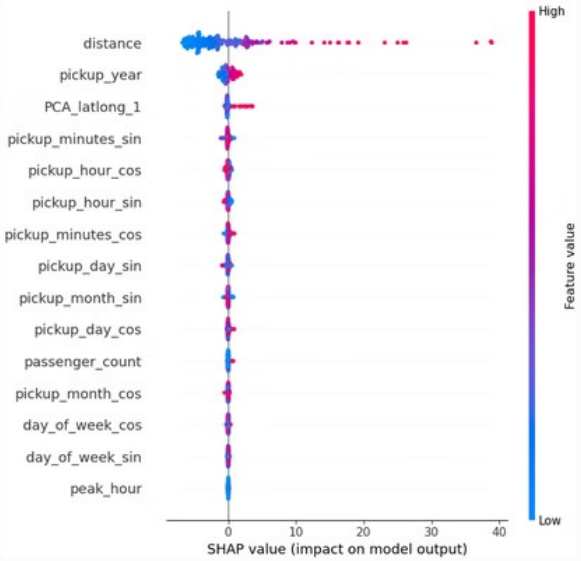
Exploración aleatoria de combinaciones de hiperparámetros para encontrar configuraciones óptimas.

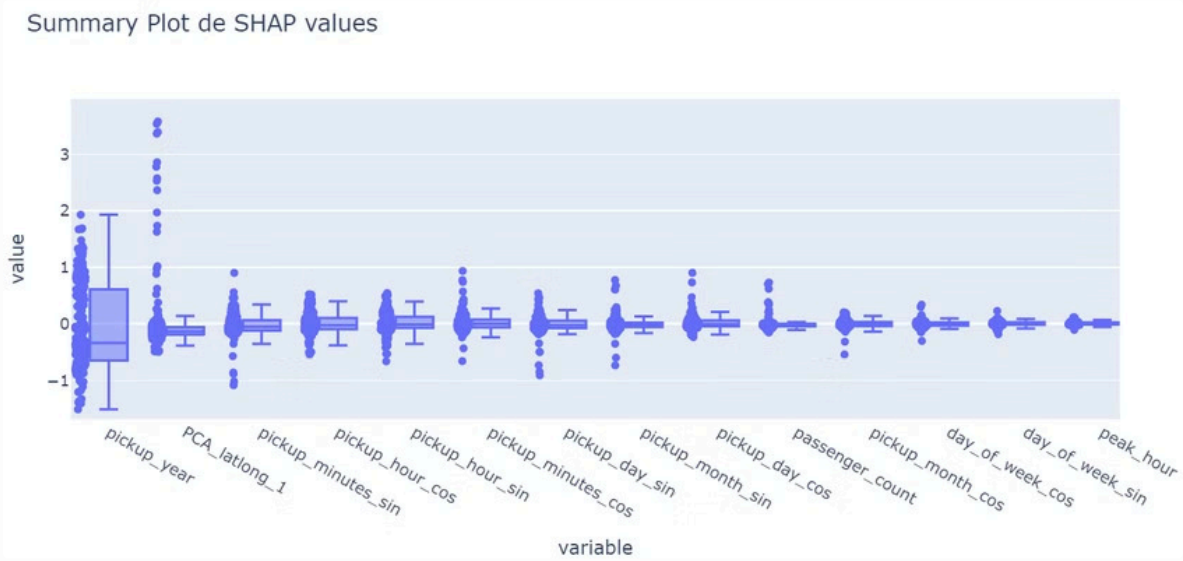
- Se crea espacio de hiperparámetros y se hacen pruebas hasta encontrar aquellos valores que ofrecen los mejores resultados.
- 5 validaciones cruzadas.



# Uso de la Librería SHAP 🚕🚕🚕🚕🚕

**SHAP** permite interpretar y explicar  cómo los modelos de machine learning toman decisiones.




La **suma** de todos los valores de SHAP y el promedio de todas las tarifas, da como resultado la predicción final para esa muestra.


# Resultados y precisión del modelo

## 81.5%

### Precisión del Modelo

El modelo final logra predecir con un 81.50% de precisión el precio de los viajes en taxi.

## 2.08 USD

### Margen de Error

El margen de error se mantiene dentro de un rango aceptable para aplicaciones prácticas.

## 100K

### Viajes Analizados

La robustez del modelo se basa en el análisis de más de 100,000 viajes de taxi.

## *

**Distancia es la variable más determinante, seguida de lejos por el día y la hora del viaje.**

# Conclusiones 🚕

## 🚕 Modelos de ML

Se destaca la comparación de la aplicación de un conjunto de modelos machine learning, y la selección de uno de ellos, en función de los resultados y  tiempo de ejecución.

Al mismo tiempo, se considera siempre necesario el entendimiento de la importancia de cada una de las variables, antes y después de la predicción con modelos.

## 🚕 Análisis de datos

En los datos utilizados no se realizó ningún tipo de estratificación o compensación. Sin embargo se hace incapié en las transformaciones y análisis de componentes principales realizados.

## 🚕 Mejoras

Se logra obtener un modelo que genera predicciones con un error medio de 2 USD.

Para un análisis más detallado, se podría combinar una eliminación de outliers y una estratificación de datos, que permita al modelo identificar los viajes por la variable más determinante, y con ello, buscar mejorar el ajuste de las predicciones.