**Proyecto de consolidación**

**Predicción precios coches**

Rafael Velasco

Andrés Flores

Manuel Rodríguez

**Business understanding:**

Tesla, Inc. es una empresa líder en innovación tecnológica y sostenible, reconocida por revolucionar la industria automotriz con vehículos eléctricos de alto rendimiento y soluciones de energía renovable.

Tesla desde 2018 dispone de 5 modelos de coche, pero solo 4 se comercializan en la UE.

Modelos:

-Model S.

-Model 3.

-Model X.

-Model Y.

En este proyecto se van a obtener datos de 3 fuentes de datos sobre distintos modelos de coches de la marca Tesla. La idea es obtener los datos mediante webscrapping y después usar esos datos para entrenar un modelo de regresión para intentar obtener el precio óptimo para un vehículo dependiendo del modelo, el color y el kilometraje.

El precio de los vehículos puede aumentar o disminuir dependiendo del kilometraje del vehículo, del color, del

Es importante que el modelo indique el mejor valor posible teniendo en cuenta las características para ajustar el precio lo mejor posible al precio de mercado del vehículo.

**Requisitos:**

Obtener los datos de necesarios de 3 fuentes distintas de datos.

El modelo tiene que obtener el precio de un vehículo dependiendo del modelo, el año de fabricación, el kilometraje y el color.

Los datos tienen que ser almacenados en una base de datos.

**Restricciones:**

Los datos tienen que obtenerse de fuentes abiertas de datos.

Los datos no pueden ser generados aleatoriamente.

**Data understanding:**

Los datos que se van a utilizar para el proyecto son:

-Modelo: Modelo del vehículo tesla.

-Precio: Precio del vehículo.

-Color: Color del vehículo.

-Kilometraje: Km hechos por el vehículo. En caso de que no aparezca dicho dato se tomará 0 de referencia.

Las páginas de las que se van a obtener los datos son:

-Página oficial de Tesla.

-Página teslahunt.io

-Página Autocasion.com

-Página Autoscout24.com

**Los algoritmos de regresión utilizados en el proyecto:**

-Decision tree regressor: Algoritmo de aprendizaje supervisado. El cual es representa condiciones sobre las características de entrada. Es fácil de entender e interpretar, pero es propenso al overfitting, si no se controla la profundidad de árbol. Se van a usar los siguientes parámetros:

·max\_depth: Maxima profundidad del árbol.

·min\_samples\_split: Cantidad mínima de ejemplares para crear un corte en el árbol.

·min\_samples\_leaf: Cantidad mínima de ejemplares para crear una hoja en el árbol.

-Random forest regressor: Algoritmo de aprendizaje supervisado. El cual es un conjunto de múltiples arboles de decisión, donde cada árbol se entrena con una muestra aleatoria de los datos y usa una selección aleatoria de características. Es más robusto, preciso y tiene a tener menor riesgo de overfitting, pero es más complejo y requiere de más potencia de cómputo, lo cual lo convierte en un modelo más lento. Se van a usar los siguientes parámetros:

·n\_estimators: Numero de árboles en el bosque.

·max\_depth: Profundidad máxima de cada árbol.

·min\_samples\_split: Número mínimo de muestras necesarias para dividir un nodo.

·min\_samples\_leaf: Número mínimo de muestras requeridas para crear una hoja.

·max\_features: Número máximo de características a considerar en cada división.

**Preparación de los datos:**

Para preparar los datos, se han obtenido todos los datos mediante webscrapping de las webs anteriormente indicadas. Y se han almacenado los datos en bruto en archivos .csv. Uno por cada página.

Posterior a la recogida de datos se procede a su correcta limpieza. Se cambian los datos de kilometraje y precio a valores enteros, puesto que al extraerlos de texto plano se encontraban con un formato y un tipo incorrecto. Posteriormente se estandariza el nombre de los modelos de vehículos obtenidos y se divide el nombre para obtener también el año de creación del vehículo.

Una vez realizadas las siguientes modificaciones se procede a estandarizar los colores de los vehículos. Puesto que dependiendo de la página los nombres pueden variar mucho entre variedades de color se ha optado por reducir los colores a los siguientes:

-Blanco.

-Negro.

-Gris.

-Rojo.

-Azul.

-Marrón.

**Modelos**:

-Decision tree regressor:

Mejor resultado:

max\_depth: None

min\_samples\_split: 2

min\_samples\_leaf: 1

Mean Squared Error: 40397427.61608545

Root Mean Squared Error: 6355.897074063224

Coeficiente de determinación: 0.9045229775671626

-Random forest regressor:

Mejor resultado:

n\_estimators': 50

'max\_depth': None

'min\_samples\_split': 2

'min\_samples\_leaf': 1

'max\_features': 'sqrt'

Mean Squared Error: 36661374.55027225

Root Mean Squared Error: 6054.863710297058

Coeficiente de determinación: 0.9133529264877934

**Conclusión:**

Los datos arrojados con ambos modelos son bastante prometedores puesto que llegamos a obtener un coeficiente de determinación de 0.91, pero aun así el root mean squared error sale relativamente alto. Aproximadamente un 15% del precio medio del vehículo. Entendemos que esa diferencia en el precio se basa en los distintos acabados interiores, extras que pueda tener el coche, colores específicos y estado de conservación general del vehículo.