Ejercicio de determinación de la calidad de carros (Clasificación múltiple)

1.- Introducción

Objetivo

Desarrollar un modelo predictivo que nos ayude a determinar la calidad de los carros, dada cierta información de entrenamiento.

Glosario

- **Modelo predictivo:** Modelo matemático probabilístico que nos ayuda a predecir o clasificar cierto evento, es decir, otorgarle una probabilidad de ocurrencia a cierta situación con base a observaciones pasadas.
- **Clasificación múltiple:** Modelo predictivo que arroja un resultado múltiple, es decir, una selección entre varias opciones, según sea el caso, es decir, la probabilidad de que se elija una de las múltiples opciones que se tienen, en este caso, un carro tenga diferentes niveles de calidad.

Planteamiento del problema

Los modelos de clasificación múltiple son utilizados para determinar dadas características de las observaciones, cual es la probabilidad de propensión a pertenecer a cierta clase. En este caso, queremos determinar la calidad de un carro con variables como son: número de puertas que tiene el carro, cuantas personas pueden entrar, el tamaño, entre otras.

Con esta información, podemos tomar una mejor decisión como clientes a la hora de comprar un carro, ya que podemos hacer un análisis de costo beneficio de lo que tiene un carro y determinar el nivel de calidad que tiene según sus características.

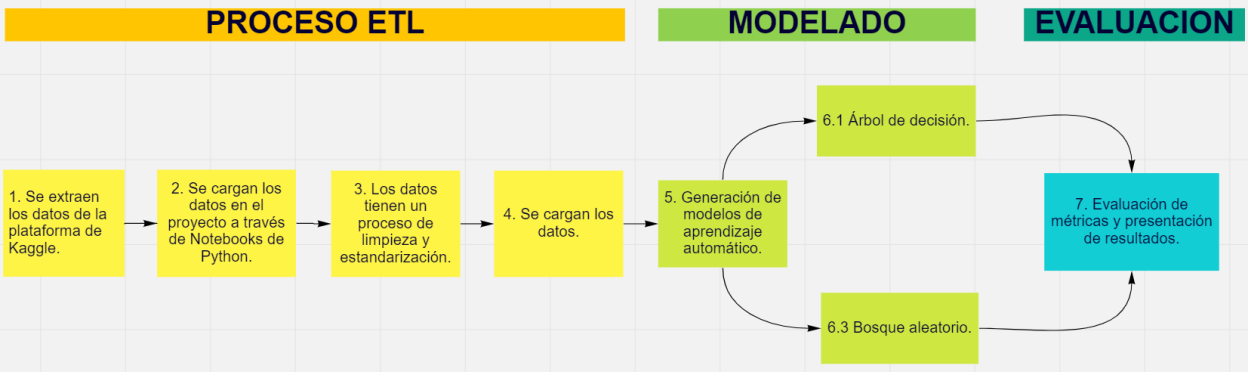
2.- Fuente de datos, flujo de datos y estructura del proyecto

2.1 Fuente y flujo de datos

Los datos para realizar este ejercicio, se extrajeron de la plataforma de Kaggel, la cual es una plataforma que ofrece una diversidad de Data sets para realizar ejercicios de ciencia de datos (El link lo encontrarán en los anexos).

El Conjunto de datos se llama "Evaluación_Carros", el cual viene en formato CSV. Las variables que vienen en el data set hacen referencia a características de los carros y la etiqueta de clasificación de calidad, según sus características.

Figura 2.1.1 Flujo de datos



La figura anterior muestra el flujo de los datos, desde su carga, modelado hasta su evaluación de modelos.

2.2 Estructura del proyecto

La estructura del proyecto se divide en 3 apartados, los cuales son:

1. **Archivos CSV:** Contiene el archivo CSV original y la salida de los datos preprocesados.
2. **Documentación:** Contiene los documentos necesarios para entender el problema y el desarrollo de este.
3. **Modelo:** Contiene los notebooks de carga y procesamiento y los 2 modelos que se probaron, así como sus métricas de evaluación.

3.- Desarrollo del modelo

3.1 Carga y Limpieza de Datos

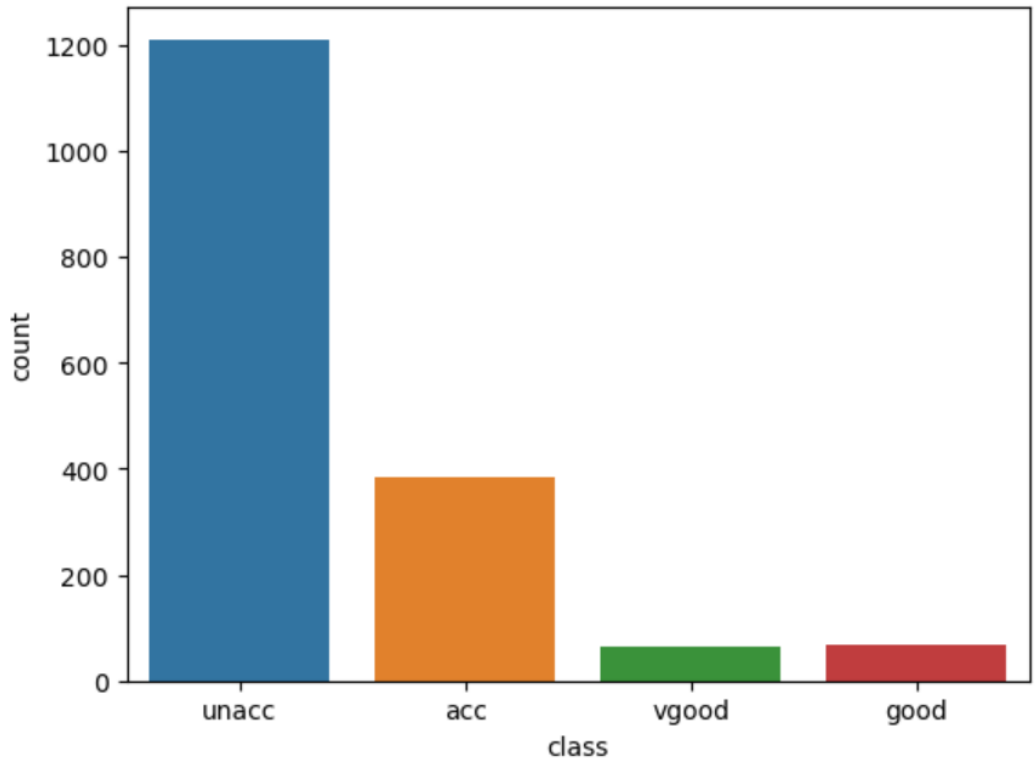
Se realizo el proceso de carga de datos como un CSV y se verificaron las dimensiones del Data Set, las cuales son de 1,728 filas (carros) y 7 columnas.

En el proceso de limpieza, se verificaron los tipos de datos y si el data set tenía valores nulos, esto para garantizar la consistencia de los datos y así el modelo pudiera arrojar mejores resultados.

3.2 Análisis exploratorio (Perfilamiento de datos)

En esta etapa se realizó un análisis exploratorio de los datos, en la cual se observa gráficamente la distribución de las clases a predecir, queda de la siguiente manera:

Figura 3.2.1. Distribución de clases de carros.



También podemos observar los números de la gráfica anterior.

Tabla 3.2.1. Tabla de distribución de clases

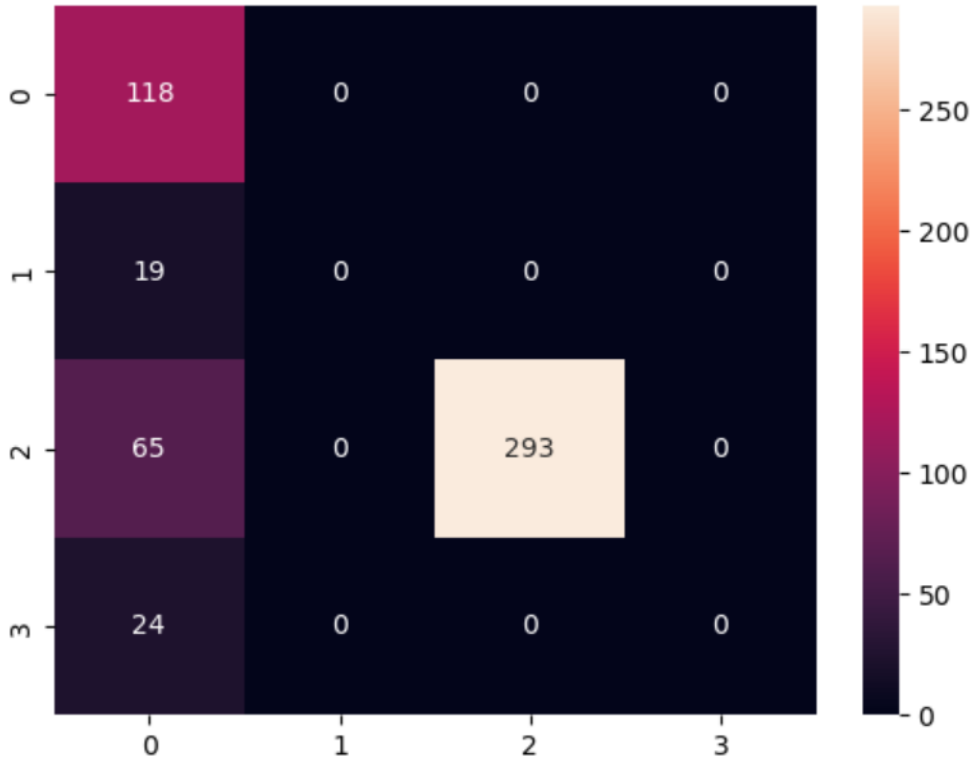
unacc	1210
acc	384
good	69
vgood	65

3.3 Modelado

En este apartado, se observarán las métricas de valuación de dos modelos diferentes, las características y resultados del modelo son las siguientes:

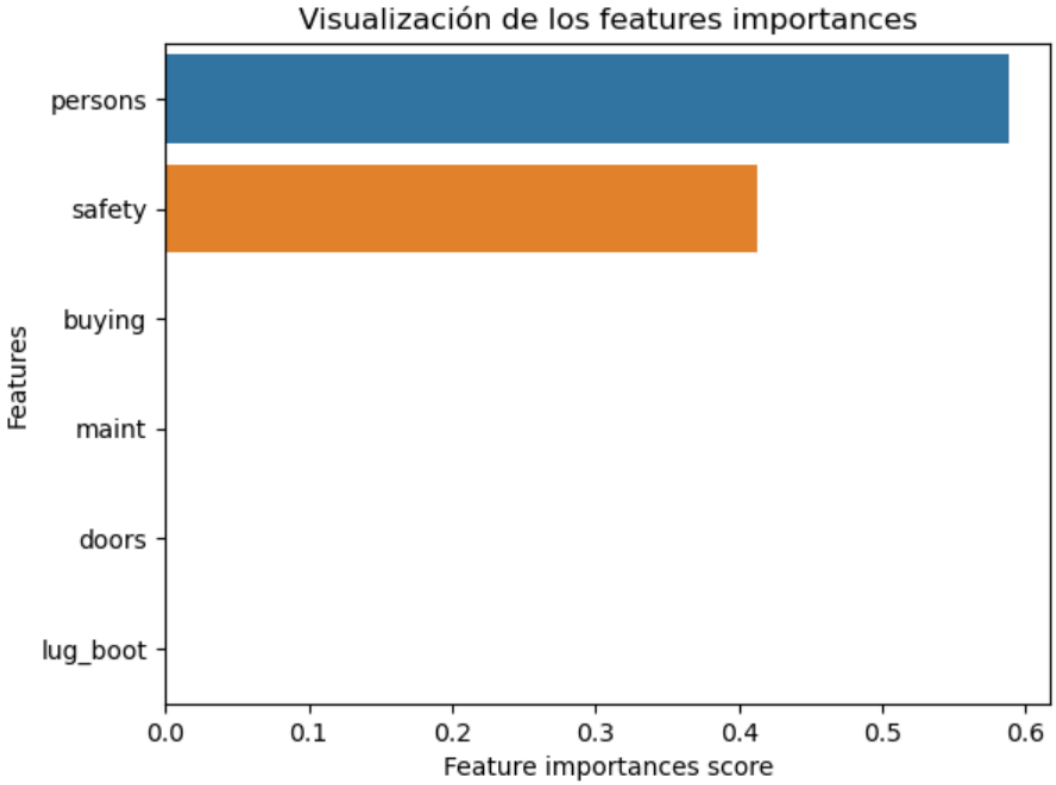
1) **Árbol de decisión:** Se generó un modelo de árbol de decisión con los datos categóricos convertidos en datos dummy Las métricas de valuación quedaron de la siguiente manera:

Figura 3.3.1.1 Matriz de confusión modelo árbol de decisión.



La matriz de confusión nos muestra que solamente hay 2 variables que determinan la predicción del modelo, para verlo más claramente, vamos a observar la importancia de las variables para el modelo:

Figura 3.3.1.2 Importancia de las variables para el modelo



Como bien mencionábamos, solamente dos variables determinan la clasificación del modelo, las cuales son: número de personas que pueden entrar y el nivel de seguridad que el carro posee.

Veremos otras métricas de valuación:

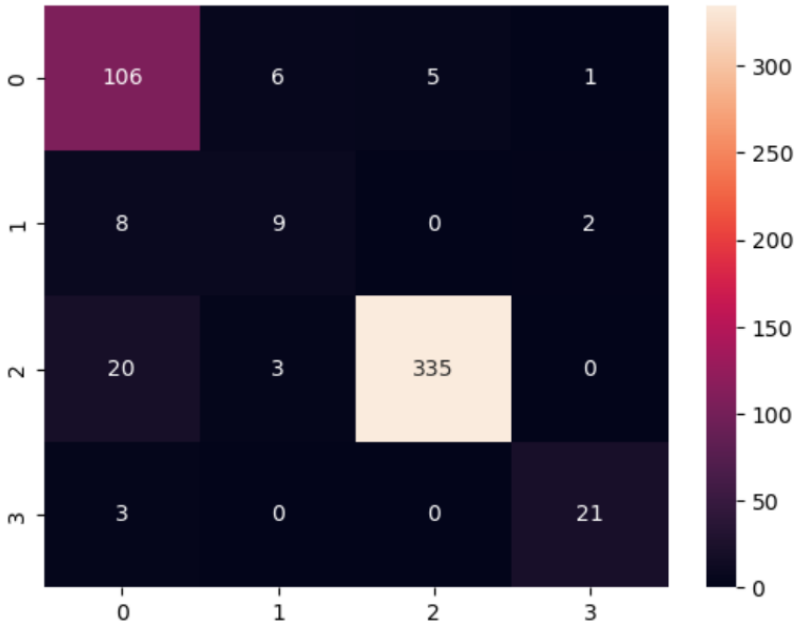
Tabla 3.3.1.1 Métricas de valuación del modelo árbol de decisión.

	precision	recall	f1-score	support
acc	0.52	1.00	0.69	118
good	0.00	0.00	0.00	19
unacc	1.00	0.82	0.90	358
vgood	0.00	0.00	0.00	24
accuracy			0.79	519
macro avg	0.38	0.45	0.40	519
weighted avg	0.81	0.79	0.78	519

Como bien mencionábamos, las métricas de valuación del modelo no son muy buenas, ya que la precisión del modelo es del 79%, pero métricas como el F1 score nos da un porcentaje de 69%.

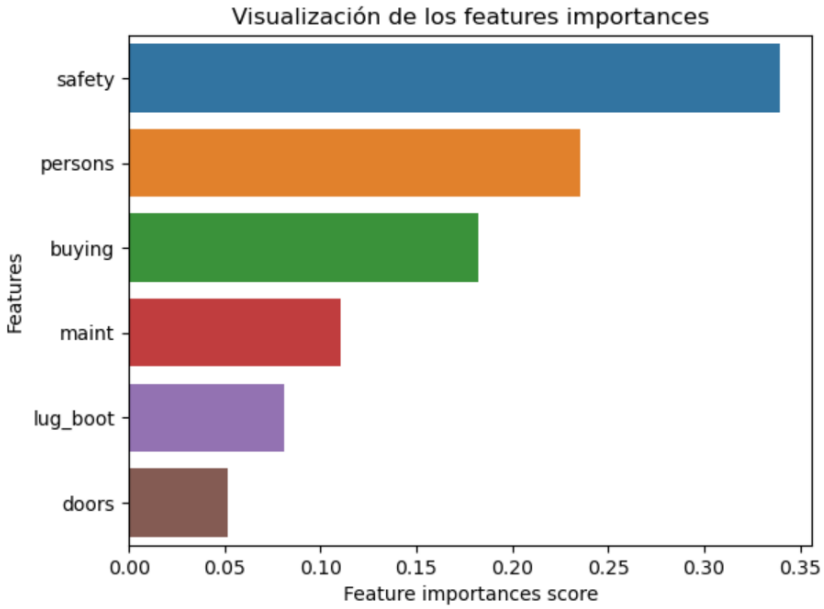
2) **Bosque aleatorio:** Se generó un modelo con un bosque aleatorio, con 2 árboles de decisión, los resultados son los siguientes:

Figura 3.3.2.1 Matriz de confusión modelo bosque aleatorio



En este modelo, hay más variables que están influyendo en el modelo, para verificar, se muestra a continuación la importancia de las variables:

Figura 3.3.2.2 Importancia de las variables modelo bosque aleatorio



En este modelo, todas las variables tienen un nivel de explicabilidad, predomina la variable de seguridad que otorgan los carros, pero todas las variables determinan la calidad de un carro.

Observaremos las métricas de medición que arroja este modelo:

Tabla 3.3.2.1 Métricas de valuación modelo bosque aleatorio

	precision	recall	f1-score	support
acc	0.77	0.90	0.83	118
good	0.50	0.47	0.49	19
unacc	0.99	0.94	0.96	358
vgood	0.88	0.88	0.88	24
accuracy			0.91	519
macro avg	0.78	0.80	0.79	519
weighted avg	0.91	0.91	0.91	519

Con este modelo, logramos llegar a una precisión del 91%. Indicadores como el F1 Score y recall, también presentan valores aceptables.

4.- Conclusiones

El modelo de bosque aleatorio, el cual se compone de dos árboles de decisión individuales, presenta una mejoría notable en la precisión de las predicciones, ya que, al ser un modelo de ensamble que nos da la maniobrabilidad de ajustar el número de árboles que queremos usar y el nivel de profundidad de los árboles, se vuelve más robusto y complejo para generarlos resultados esperados.

5.- Anexos

- Repositorio en Git Hub:
  https://github.com/luis151294/Calidad_de_Carros

- Data Set en Kaggle:
  https://www.kaggle.com/datasets/elikplim/car-evaluation-data-set