**GUIA DE USUARIO Y DESCRIPCIÓN DEL PROYECTO**

**“PREDICCIÓN DE PRECIOS DE COCHES DEPORTIVOS”**

**DESCRIPCIÓN DEL PROCESO:**

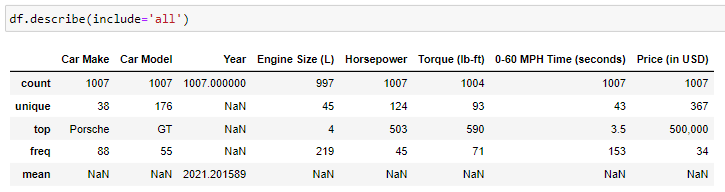
El proceso de realización del proyecto ha constado de dos partes diferenciadas:

1. Modelización y análisis de la base de datos
2. Creación de la aplicación en streamlit

**Modelización y análisis de la base de datos**

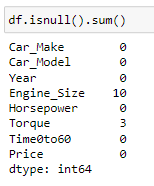
Primero de todo, cargamos el dataset del csv descargado en Kaggel (<https://www.kaggle.com/datasets/rkiattisak/sports-car-prices-dataset>)

Una vez cargado, revisando las primeras y últimas líneas y vemos que el dataset se compone de 1007 filas x 8 columnas, es decir 1007 datos y 8 características. Como podemos ver a continuación la distribución de las mismas:



Con la instrucción de describe observamos que hay dos columnas “Engine Size(L)” y “Torque (lb-ft)” no tienen la totalidad de los datos.

Si con el describe no vemos claro que hay parámetros que no tienen la totalidad de los datos podemos buscar si hay valores vacíos en alguna variable. Para nuestro análisis vemos que tanto “Engine Size(L)” y “Torque (lb-ft) contienen 10 y 3 casillas vacías respectivamente y prescindiremos de ellas:

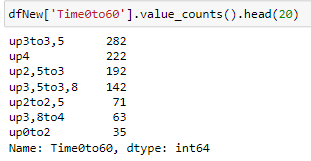


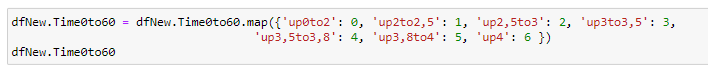
Cuando ya tenemos el primer contacto con el dataset y vemos que tipo de variables son y como están distribuidas, hacemos el “Feature Engineering” transformando las columnas que tenemos para que puedan ser analizadas. Por lo que, las que tienen una amplitud de valores las agruparemos por intervalos.

Tenemos que “Time0to60” y “Horsepower” son variables compuestas de un rango amplio de valores y las agruparemos para poder analizarlas mejor:

Time0to60:

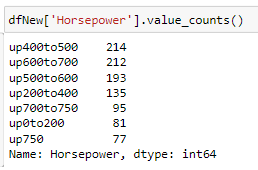
Esta variable nos indica el tiempo de aceleración del coche de 0 a 60 segundos, este tiempo de aceleración puede tardar de 1.85 to 6.5 en conseguirse dependiendo del coche. De todas formas, consideraremos los siguientes intervalos que luego los categorizaremos numéricamente del 0 (de 0 a 2 seg.) al 6 (más de 4 seg.)





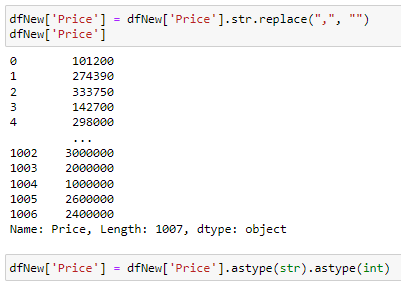
Horsepower:

Esta variable nos indica los caballos de potencia que tiene el coche. Los valores que esta variable toma es de 181 to 1479. De la misma forma que la variable “Time0to60”, con la variable “Horsepower” consideraremos los intervalos que se observan en el siguiente pantallazo de código y además categorizaremos numéricamente del 0 (de 0 a 200) al 6 (más de 750)





A variable precio le hacemos también unas modificaciones para poder tener los datos en un formato que se pueda analizar:



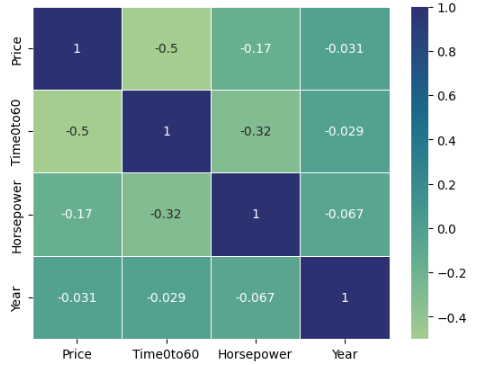
Después de realizar todos los cambios y pasar a variables numéricas los parámetros del Tiempo de aceleración de 0 a 60 segundos, de los caballos de potencia y del precio tenemos que las variables de Marca de coche y modelo habrá que pasarlas a formato numérico para ser analizadas.

La marca y coche del modelo las modelizaremos a través de un “get\_dummies” esto lo que nos creará será una columna para cada marca y cada modelo de coche indicando con un 0 o 1 si el coche y marca pertenecen a esa categoría este punto hace que el dataset de tener 6 columnas pasemos a tener 218. Usamos la instrucción de “get\_dummies” ya que el comando se utiliza para la manipulación de datos y lo que hace es convertir datos categóricos en variables indicadoras. Es un punto importante ya que en el momento de calcular y a pasar los datos que el usuario introduce, el vector de datos que pasamos al modelo debe tener tantas columnas como columnas haya en el modelo calculado. Esto lo solucionaremos mediante un bucle for el cual mirará que marca y modelo determinado le ha pasado el usuario y con sólo esos dos parámetros aparecerán con el 1 correspondiente en sus variables y al resto de columnas/parámetros les pasaremos el 0.

Con toda esta información, podemos concluir que las variables analizables serán Year, Horsepower, Time0to60, Price, Car\_Make y Car\_Model con estas dos últimas parametrizadas por marca y modelo de coche.

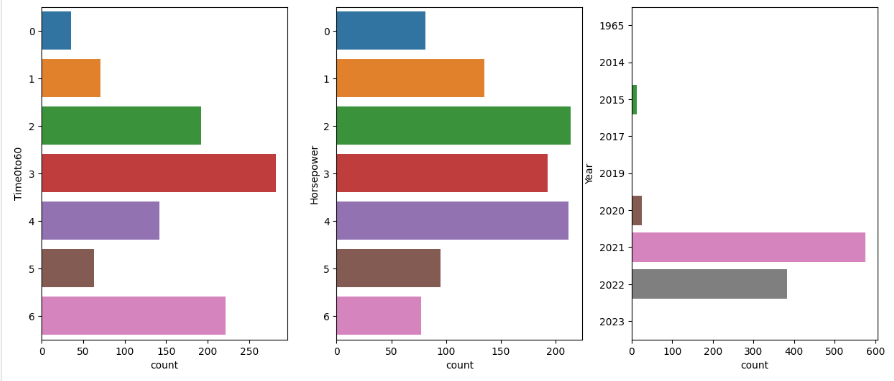
Ahora ya tenemos todas las variables en el formato que necesitamos para obtener una predicción del modelo pero antes de ello analizaremos mediante gráficos los datos que tenemos para ver si hay alguna relación entre ellos y si hay algo a destacar.

En el gráfico de calor (heatmap) analizamos las variables que no tienen que ver con la marca y modelo del coche y vemos que no tienen ningún tipo de correlación unas con otras, ni negativa ni positiva destacable.



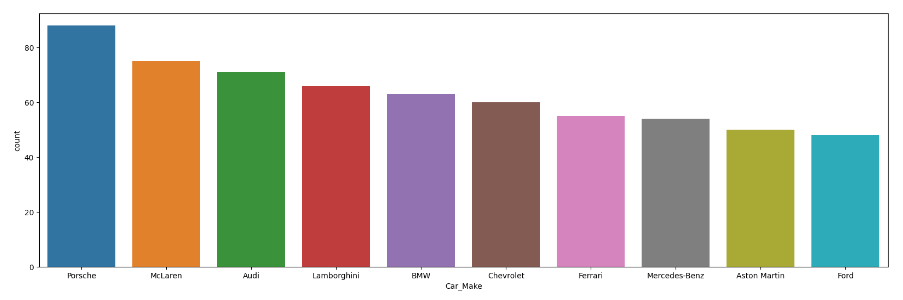
Si miramos el grafico de calor (heatmap) con las variables de marca y modelo parametrizadas debido a su gran tamaño no nos aporta ningún tipo de información es por eso que miraremos las variables dos a dos o por grupos.

Realizamos varios análisis a través diferentes tipos de gráficos para ver si podemos obtener alguna información de cómo se comportan las variables:



En el segundo grafico podemos observar que la mayoría de datos que tenemos se concentran en las categorías que se encuentran entre la 2 y la 4, es decir, la mayoría de datos que tenemos tienen de 400 a 700 caballos.

Sobre las marcas de los coches destacamos el Top 10 que tenemos mayor número de datos:



Tenemos el Top 10 de las marcas que más datos tenemos, indicar que en la aplicación incluiré únicamente las marcas que tienen más de 10 datos ya que considero que si no hay datos suficientes los resultados no son significativos.

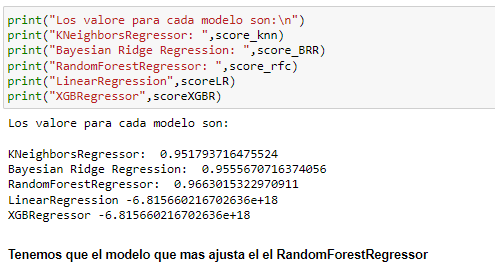
Para el entrenamiento de datos vamos a hacer las partes correspondientes para tener la X\_train & Y train como el 80% de datos y para el X\_test i el Y\_test usaremos el 20% restante. Así de ese 20% analizaremos el porcentaje de acierto y así veremos cuál es el modelo que más se ajusta a los datos que tenemos. Por lo tanto tenemos que:

X = todas las variables excepto el precio

Y = precio (característica a predecir)

El tamaño del entrenamiento será del 80% del dataset.

Debido a que es un modelo de regresión múltiple, ya que tenemos que predecir un precio a partir de varias variables usaremos varios los modelos de regresión para ver cual se ajusta mejor a nuestro modelo de datos.



De todos los modelos que hemos testado el que mejor aproxima los datos del modelo para la predicción del precio de coches deportivos es el Random Forest Regressor.

Para el modelo de Random Forest Regressor nos da un valor de un 96.6% de aproximación. Consideramos que los valores mayores de 95% son bastante buenos.

Para finalizar con la parte de análisis y modelización lo que haremos será fijar el modelo de predicción para agilizar los cálculos cuando haya que predecir el precio de un coche deportivo. Para realizar este paso usaremos la instrucción “pickle”



Con esta instrucción crearemos un archivo con extensión pkl (en nuestro caso model.pkl) que será el archivo que será llamado por la aplicación para calcular.

**Creación de la aplicación en Streamlit**

Esta aplicación está realizada para que cualquier usuario que esté interesado en la compra de un coche deportivo pueda obtener una predicción del coste que tendrá el coche con las características señaladas que se solicitan. Para hacer dicha predicción es necesario introducir ciertos parámetros (características) del coche para poder ajustar el precio orientativo.

[Una vez tengo el modelo de predicción escogido, empiezo a ver que variables son necesarias para realizar la predicción. Las variables con las que he decidido realizar el modelo son básicamente 5, Marca y modelo del coche, año del coche, tiempo de aceleración de 0 a 60 segundos y los caballos de potencia.]

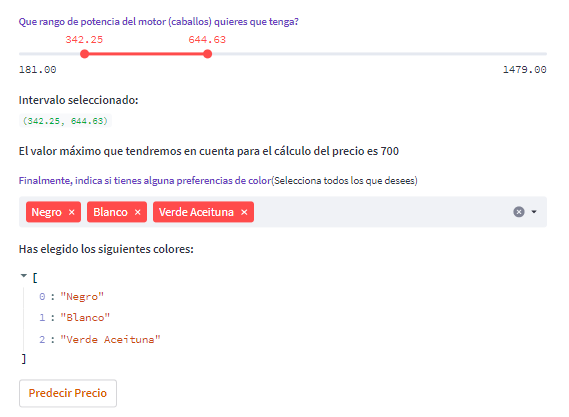
Primero de todo, cuando ejecutamos el programa y accedemos a él, podemos observar que la página se divide en dos columnas diferenciadas para tener una visión más compacta, más adecuada a la petición de los datos y visualmente más intuitiva.



En el primer bloque, encontramos en la parte de arriba a la izquierda un desplegable con la marca de coche a elegir. Luego, dependiendo de esta elección veremos que la foto del coche se identifica con la opción que hemos elegido de la marca. [Esto hace que visualmente más atractivo y te enseña que tipo de coche has elegido (una imagen de uno de los posibles modelos a seleccionar).]

En el siguiente bloque tenemos que, la parte de la izquierda consta de dos selecciones diferentes, por un lado hay que seleccionar el año que queremos del coche y luego podemos seleccionar el tiempo de aceleración de 0 a 60 segundos que queremos que el coche tenga. Esta opción se compone de 7 posibles intervalos. En la parte de la derecha encontramos los posibles modelos que podemos seleccionar de la marca indicada anteriormente. En este caso para Maserati tenemos dos pero dependiendo de la marca se pueden seleccionar tantos como modelos tenemos en el dataset.

Una vez hemos indicado las características de marca, modelo, año del coche y tiempo de aceleración de 0 a 60 segundos, solamente nos quedará indicar la potencia del motor que queremos y con esta última variable podremos realizar la predicción del precio del coche deportivo que queremos.



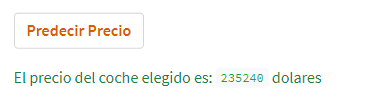
El usuario, tiene la posibilidad de seleccionar un intervalo en el que quiere que se encuentre la potencia del motor de su coche aunque nosotros nos quedaremos con el valor máximo como valor de referencia y debido a que hemos clasificado la variable por intervalos será este valor que pertenece a un intervalo concreto el que le pasaremos al modelo para evaluar. Se ha elegido el mínimo Horsepower de 181 y el máximo de 1500 ya que son los valores posibles en el dataset.

Como el modelo cuenta con unos intervalos para esta variable tendremos que calcular en que intervalo de potencia se encuentra el valor elegido por el usuario y es por este motivo que hemos añadido la frase “El valor máximo que tendremos en cuenta para el cálculo del precio es:” a modo informativo para que el usuario sepa la potencia de motor (caballos) con la que puede contar.

Cuando ya tenemos todas las variables necesarias le consultamos al usuario para finalizar mediante un multi-selector que nos deja elegir las diferentes opciones que color quiere que tenga el coche. [Este dato no es necesario pero me parecía una característica a tener en cuenta.]

[Una vez obtenidos todos los datos tenemos que crear el vector que contenga la información correcta de todos los datos introducidos y cumplimentar los datos hasta obtener un array que contenga la totalidad de variables que el modelo necesita para realizar la predicción. Es por este motivo por el que tenemos que crear un bucle que nos incorpora todas las variables nulas y nos indica con un 1 cuando la marca y el modelo coinciden.]

Y con todo esto ya podemos hacer la predicción del modelo. Una vez ejecutado veremos el precio que nos predice.



Con las características anteriores (por ejemplo) se tiene que el precio que predice es de 235.240$

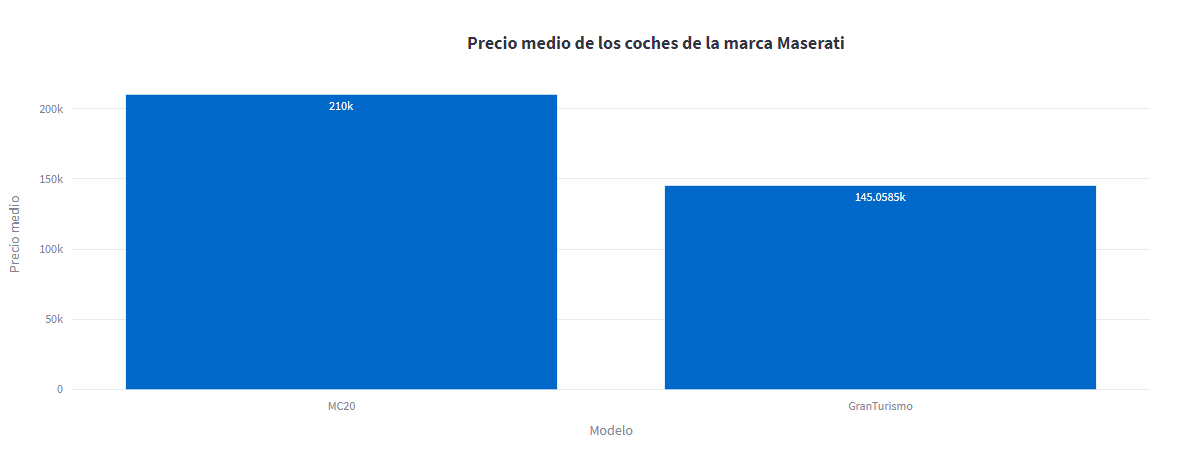
Una vez obtenido el precio del coche del que se ha hecho la predicción con las características concretas seleccionadas de éste, encontramos a continuación unos gráficos que nos enseñan varios datos de la marca y modelo escogido.

En todos los gráficos tenemos los mismos comandos en la parte superior derecha que nos pueden dar más concreción o hacer zoom en una parte del gráfico como se puede ver en el siguiente recuadro:

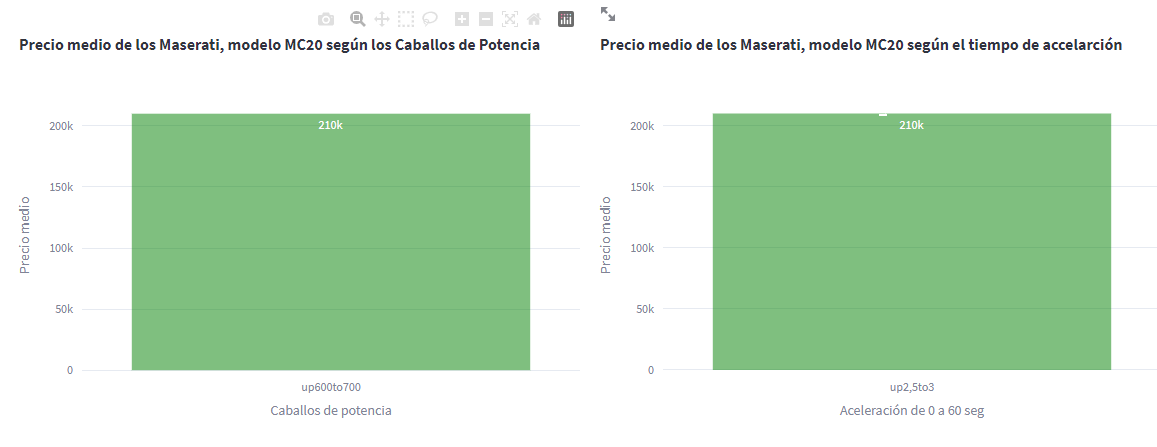


Todos estos iconos nos permiten interaccionar con el grafico que tenemos seleccionado y podríamos, si nos interesa, desde hacer una foto – icono de la izquierda – a ampliarlo a pantalla completa – icono de la derecha – donde se encuentran las dobles flechas volteadas.

En esta batería de gráficos primero de todo observamos el Precio medio de los modelos que pertenecen a la marca seleccionada. En este caso, Maserati: Como podemos ver en el gráfico el modelo MC20 con un precio medio de 210.000 dólares y el modelo Gran Turismo con un precio medio 145.058 dólares.



Luego, a continuación obtenemos información de los caballos de potencia y de la aceleración de 0 a 60 segundos de la encontramos dos gráficos que teniendo en cuenta la Marca y modelo del coche escogido. En este ejemplo tenemos que para Maserati modelo MC20, todos los datos que tenemos en el dataset es de tenemos son de coches es con un máximo de 700 caballos y una aceleración de 2.5 a 3 segundos.



De la misma forma que en el resto de gráficos en estos dos disponemos de las opciones de optimización a pantalla completa, foto y ampliación.

Y, el último grafico que podemos ver en la aplicación es un gráfico “estático”, es decir, en el que la información no cambia. Siempre nos muestra los mismos datos independientemente de las variables elegidas anteriormente. Nos muestra cuales son las marcas con precio mayor de 1M de dólares y nos indica el precio medio de los coches que comercializa cada una de las marcas de coche.

