

# Proyecto Final

## ❖ Abstract

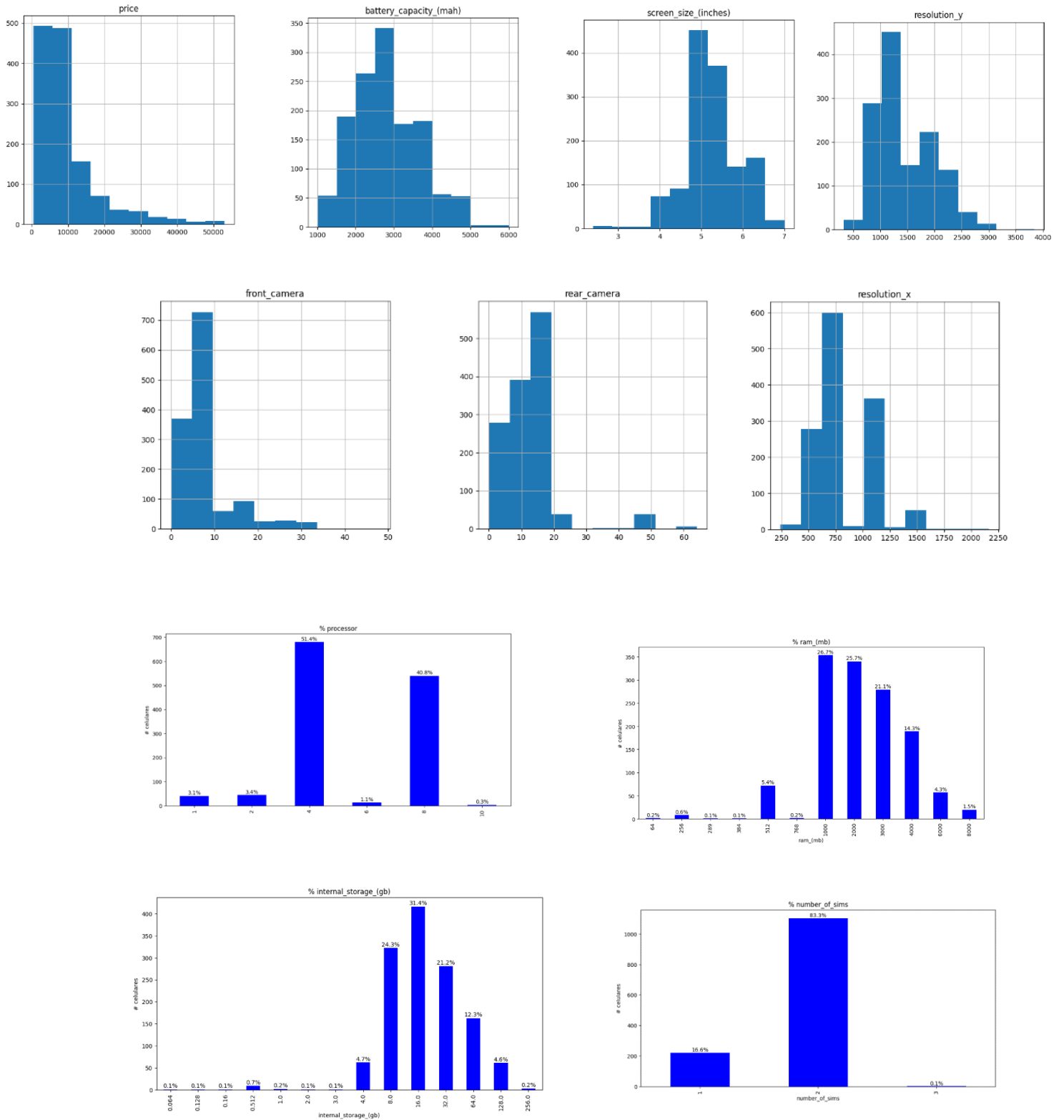
- **Objetivo** - analizar las relaciones y los patrones dentro del conjunto de datos para obtener información sobre los factores que pueden influir en los precios de los teléfonos móviles.
- **Contexto comercial** - Una muy reconocida empresa de venta de celulares me solicita crear un modelo que prediga (basándose en ciertos datos) si la marca de determinado celular es tope de gama o no. Esto podría ser beneficioso ya que permite analizar las nuevas marcas de teléfonos, posibles competencias, que se están creando últimamente, como por ejemplo Nothing Phone.
- **Contexto analítico** - exploraré las correlaciones entre las diferentes variables, podemos identificar los determinantes clave que impactan en la fijación de precios de los teléfonos inteligentes.
  - **Hipótesis** - basándome en mi experiencia, lo que más espero que tenga impacto en los precios será la marca, la gama y el procesador.

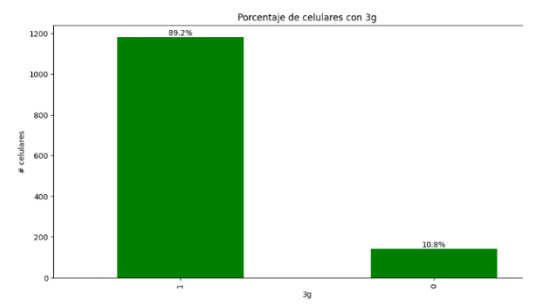
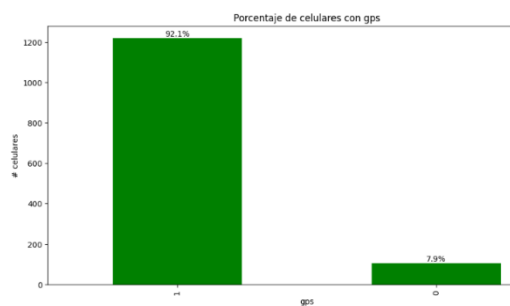
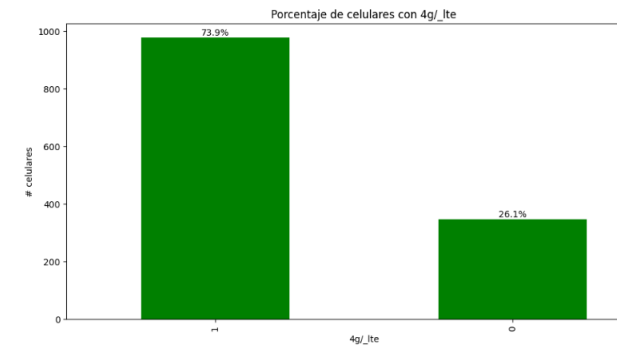
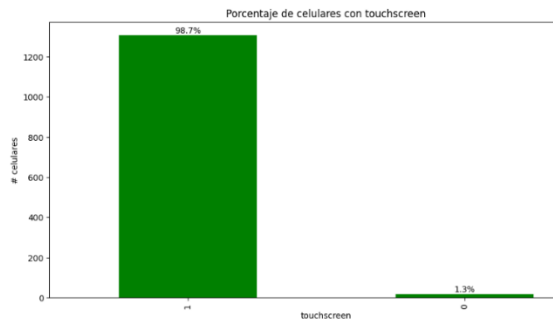
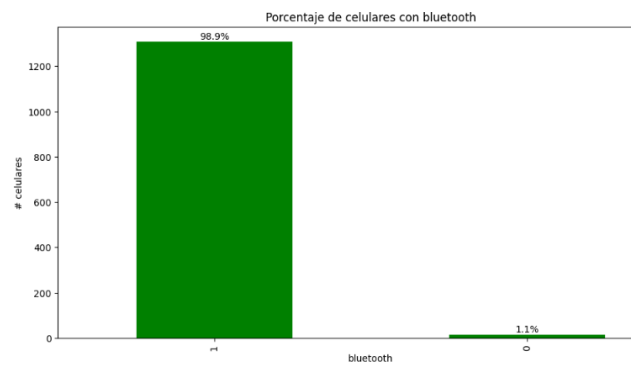
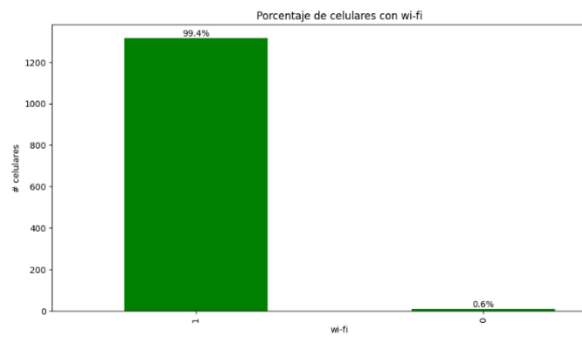
## ❖ Adquisición de datos

Este estudio examina un conjunto de datos extraídos de publicaciones de ventas en Amazon, contiene información sobre los precios de los teléfonos móviles y diversas características (como el nombre, marca, modelo, capacidad de la batería, etc).

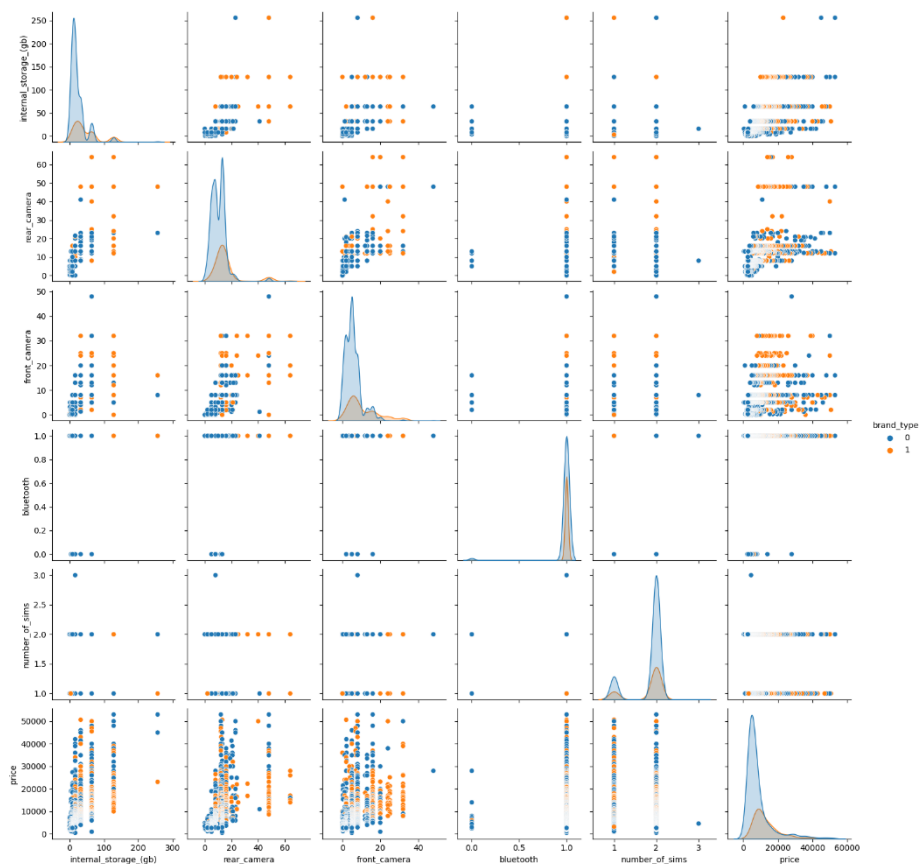
## ❖ EDA

- **Análisis univariado** – las variables fueron separadas según sean: dos valores, algunos valores o muchos valores para hacer los plots.





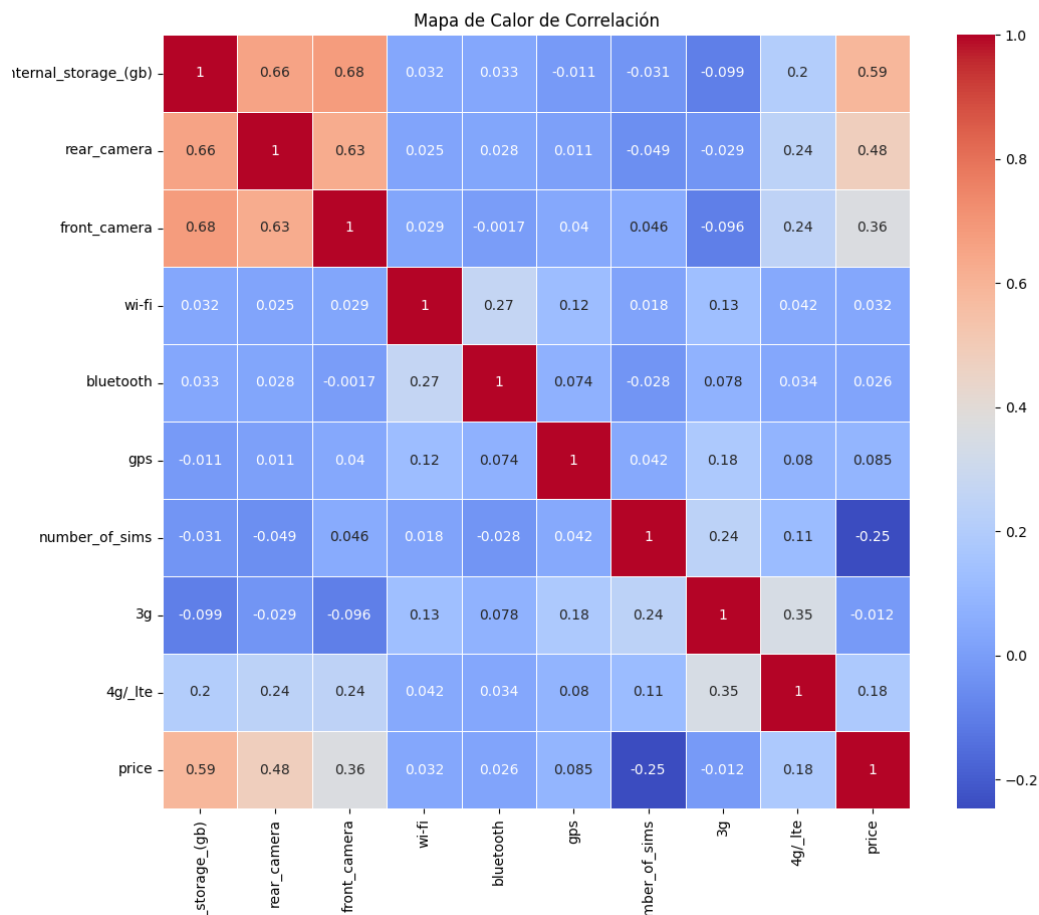
○ ***Análisis bivariado*** - Analizado mediante un paiplot



- **Análisis multivariado** - Estudio abarcando gran cantidad de variables simultáneamente

■ Correlaciones más fuertes:

- front\_camera y internal\_storage (0.68)
- rear\_camera y internal\_storage (0.66)
- front\_camera y rear\_camera (0.63)



## ❖ Feature Engineer

Se agregaron las columnas "brand\_type" la cual tiene valores: "Famosa" o "No famosa" según la marca, tomado de un dato extra-estadístico sacado de un estudio realizado por Omdia en el 2Q2022, dividiendo las marcas entre "reconocidas mundialmente" y "pequeñas".

Además, he agregado la columna “gama” de valores: “Alta” o “Baja”, la cual toma el promedio de precio de la marca, si el valor del celular es mayor a este lo considera alto será de gama alta y si está por debajo, gama baja.

Por último, agregué la columna “top”, donde separe en 3 grupos todos los celulares según su precio. Los tamaños de los grupos pueden ajustarse.

## ❖ Entrenamiento de modelos

- **Split** – el Dataset fue separado en un 80% para entrenar al modelo y 20% para probarlo.
- **Transformaciones** – se transformaron los valores binomiales a 0 o 1 y se utilizó Label Encoder para poder transformar los sistemas operativos a números.
- **Entrenamiento** – se entrenaron los modelos:
  - Decision Tree Classifier, utilizando las selecciones de variables Forward y Floating.
  - Support Vector Machine, buscando diferentes valores del parámetro c.
  - Random Forest, utilizando LeaveOneOutCrossValidation.

## ❖ Conclusión

Se utilizó la medida F1 Score ya que el Dataset representaba un desbalance de 73.3% a 26.7%, a favor de las marcas “desconocidas”.

Los resultados concluyeron que el modelo Random Forest con un  $c = 0.1$  posee un mayor nivel de accuracy, aproximadamente acertando en un 78% de los casos.