# Capstone Project

# Mobile Price Range Prediction (Supervised ML- Classification)

**Individual Project**
**Tanmaya Kumar Pattanaik**

# Introduction

In the competitive mobile phone market companies want to understand sales data of mobile phones and factors which drive the prices.

The objective is to find out some relation between features of a mobile phone(eg:- RAM, Internal Memory, etc.) and its selling price.

# Mobile Price range Data

In this project, we are going to explore and analyze a dataset which contains specifications of two thousand mobile phones and try to predict optimum price ranges for a list of mobile phones in the market by applying various machine learning algorithms such as logistic regression, decision tree, random forest and k-nearest neighbors(KNN).

# Column Information

- **Battery_power-** Total energy a battery can store in one time measured in mAh
- **Blue** - Has bluetooth or not
- **Clock_speed** - speed at which microprocessor executes instructions
- **Dual_sim** - Has dual sim support or not
- **Fc** - Front Camera mega pixels
- **Four_g** - Has 4G or not
- **Int_memory** - Internal Memory in Gigabytes
- **M_dep** - Mobile Depth in cm
- **Mobile_wt** - Weight of mobile phone
- **N_cores** - Number of cores of processor
- **Pc** - Primary Camera mega pixels
- **Px_height** - Pixel Resolution Height
- **Px_width** - Pixel Resolution Width

# Column Information(Contd.)

- **Ram** - Random Access Memory in Mega Bytes
- **Sc_h** - Screen Height of mobile in cm
- **Sc_w** - Screen Width of mobile in cm
- **Talk_time** - longest time that a single battery charge will last when you are
- **Three_g** - Has 3G or not
- **Touch_screen** - Has touch screen or not
- **Wifi** - Has wifi or not
- **Price_range** - This is the target variable with value of 0(low cost), 1(medium cost), 2(high cost) and 3(very high cost).

# Required Packages

- Numpy
- Pandas
- Matplotlib
- Seaborn
- Sklearn Packages:

1. Train_test_split
2. Metrics
3. Logisctic regression
4. Decision tree classifier
5. Random forest classifier
6. Kneighbours classifier
7. Classification report
8. Confusion matrix
9. Accuracy score

# Data exploration

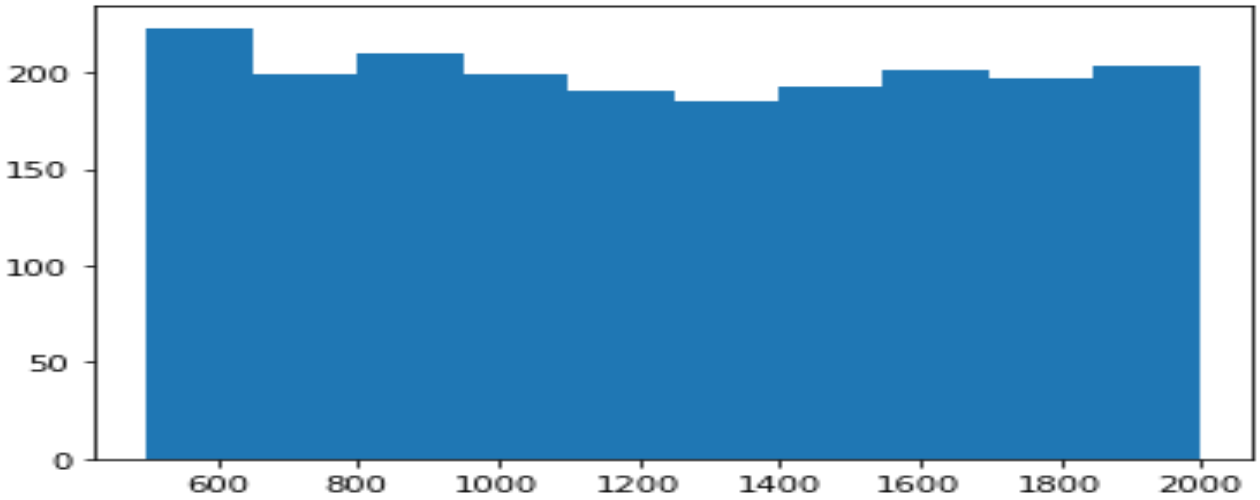
- There are 2000 rows and 21 columns in our dataset.
- The last attribute i.e., price_range_column is a target variable.
- The dtype of all attributes are in int or float type.
- There are no Nan/Null values in the dataset.
- We defined our target column as “Y” and rest of the data which are used as inputs as “X”.
- There are four price ranges as target, so we did multi-class classification in our project.
- Also our dataset is balanced with 25% being each share of price range.

# Exploratory data analysis

Let's deep more into the data and understand the dataset by exploring all columns one by one with the help of visualization. It will help us in understanding and building models.

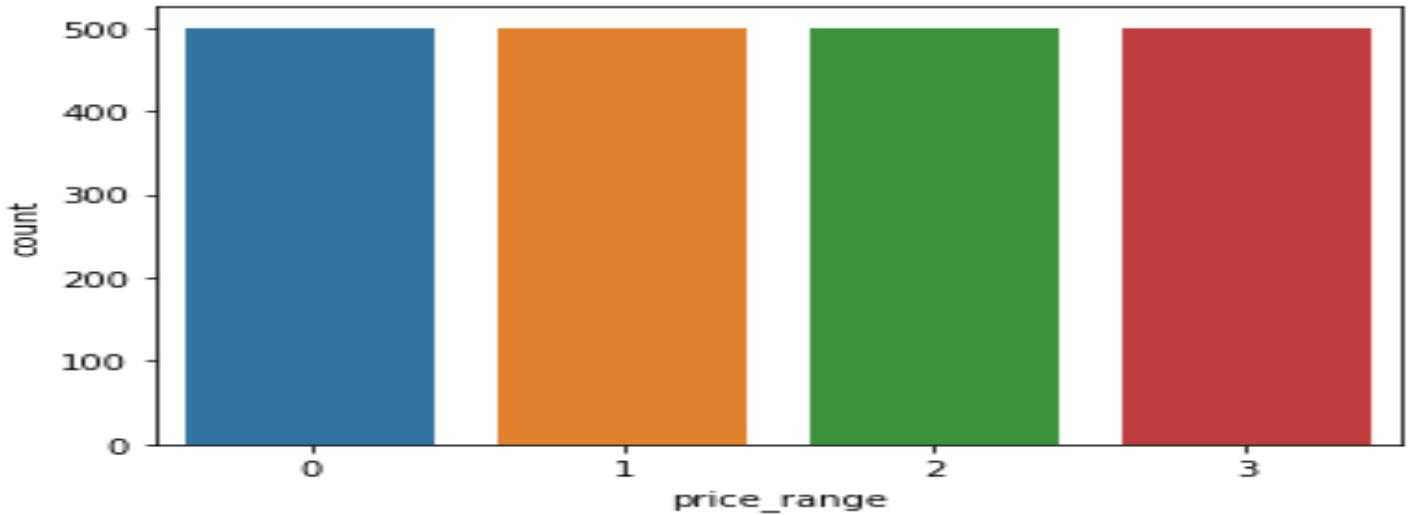
We will explore each columns and compare with our dependent variable i.e., Price range and see how they are related to each other. It will help us in analyzing which variables are the main factors in pricing of mobiles.

# Battery Power



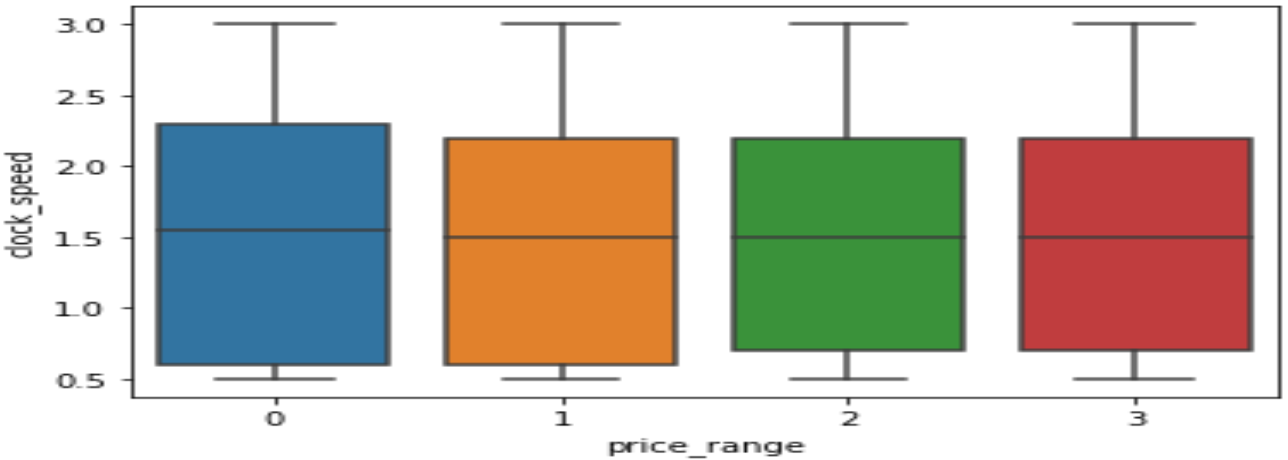

Low power batteries are slightly more in count.

# Price Range



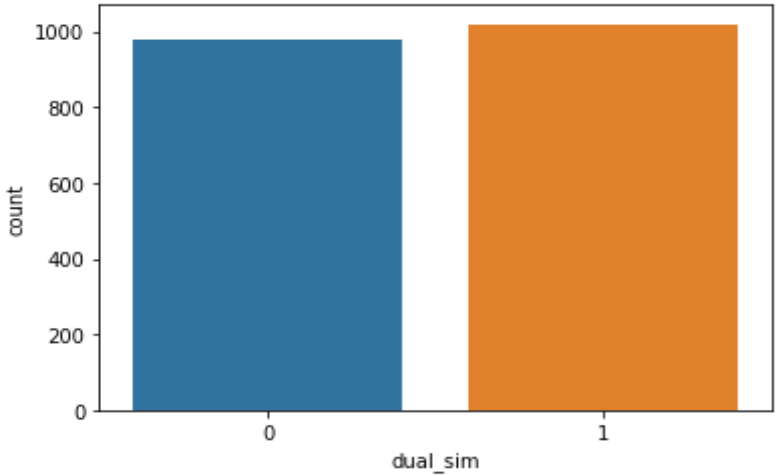

Its quite a uniform data. Data is split equally across all ranges.

# Clock Speed

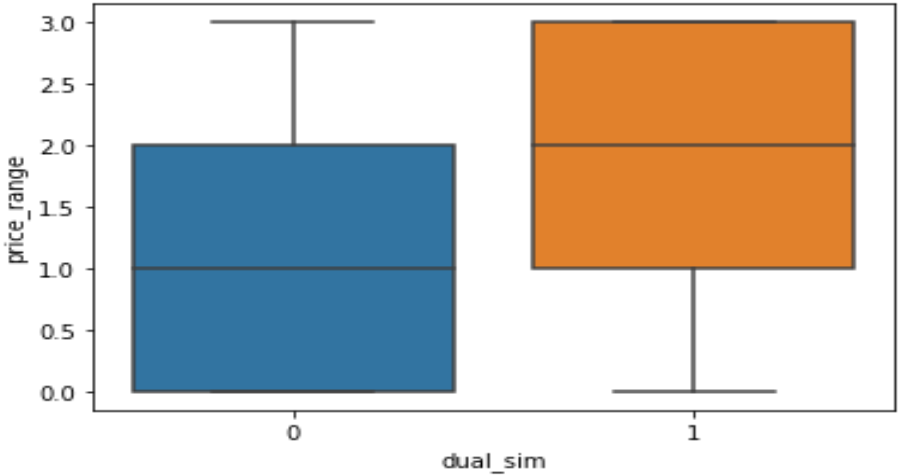



Variance of clock speed is slightly more for mobiles in Category '0'

# Dual Sim

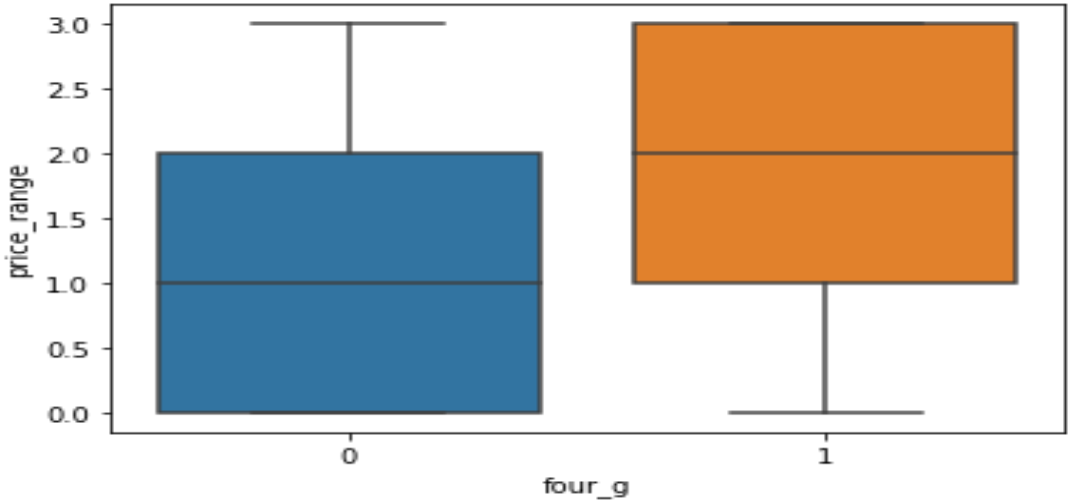



Slightly a more number of phones have dual sim



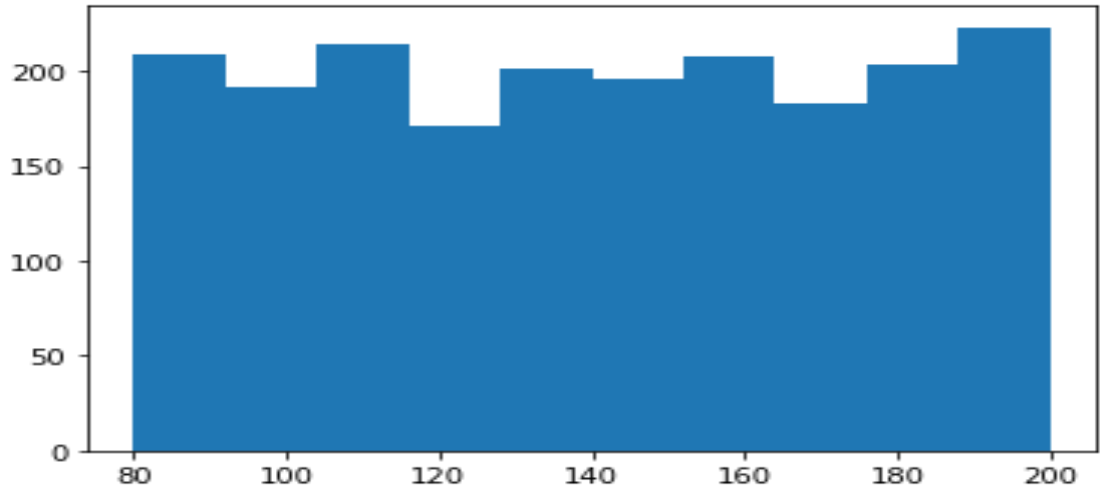

Price Range of dual sim phones are considerably higher.

# Four G



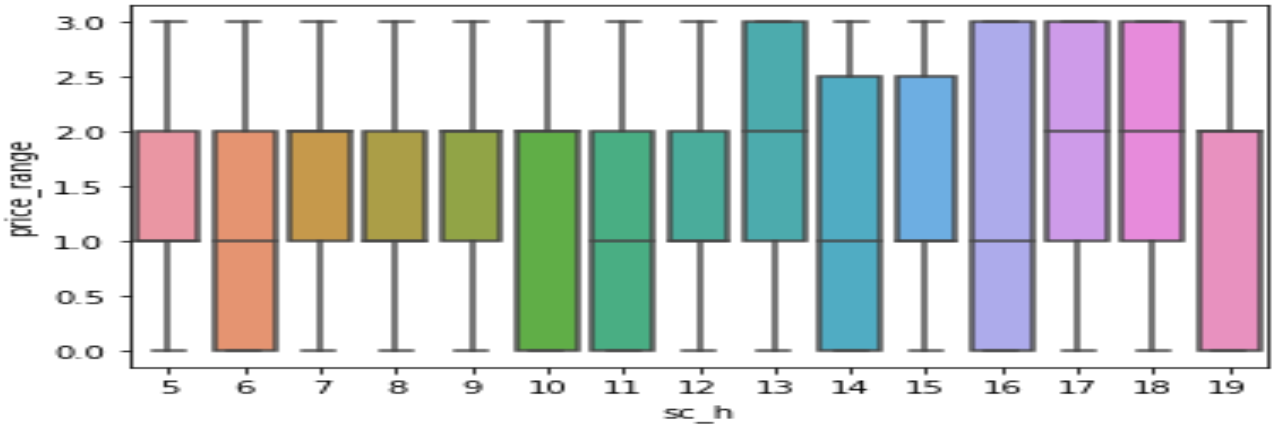

Price Range of 4G phones are considerably higher.

# Mobile Weight



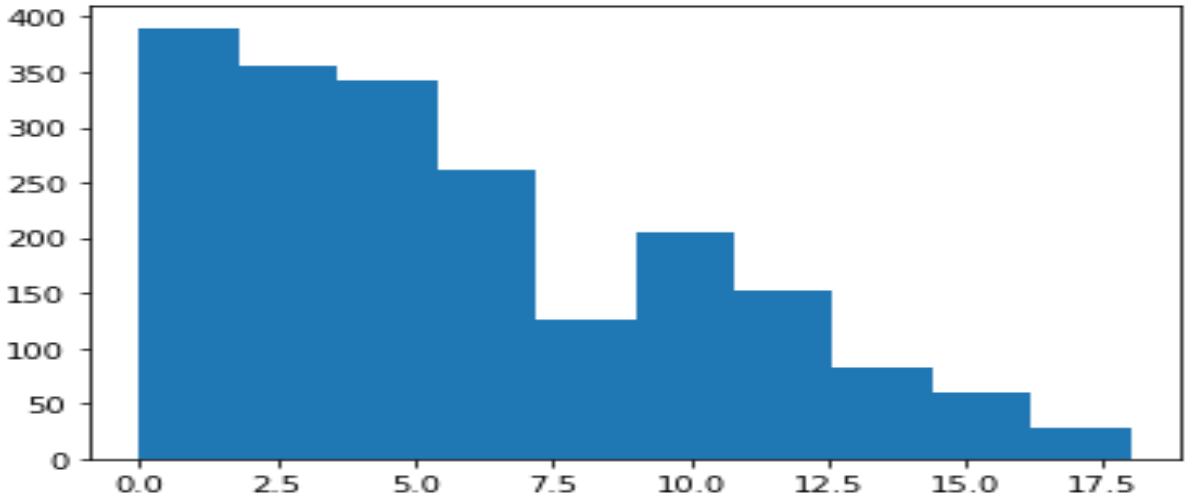

Almost evenly spread across data set

# Screen Height of mobile in cm



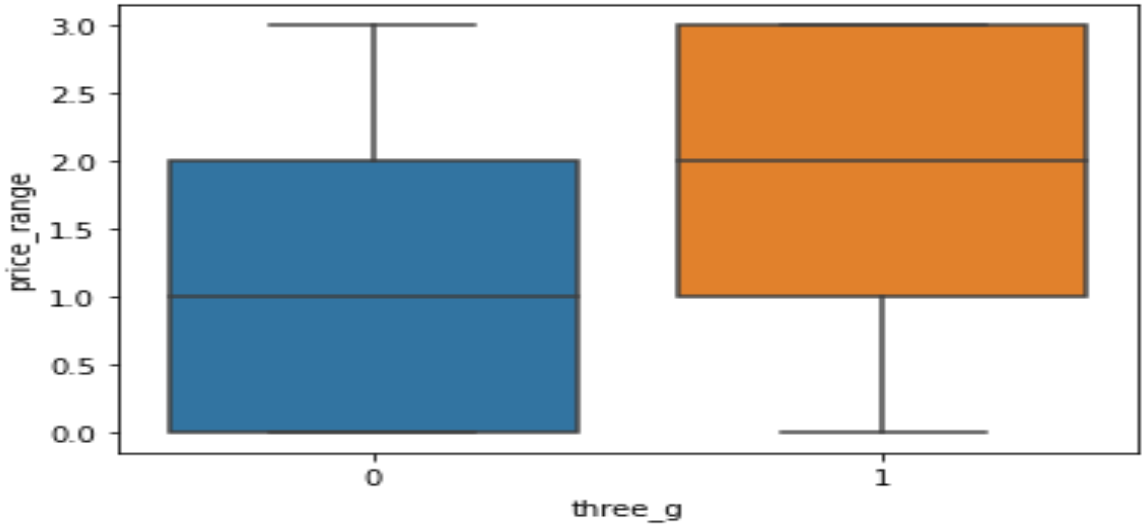

Some screen sizes are in high price range

# Screen Width of mobile in cm



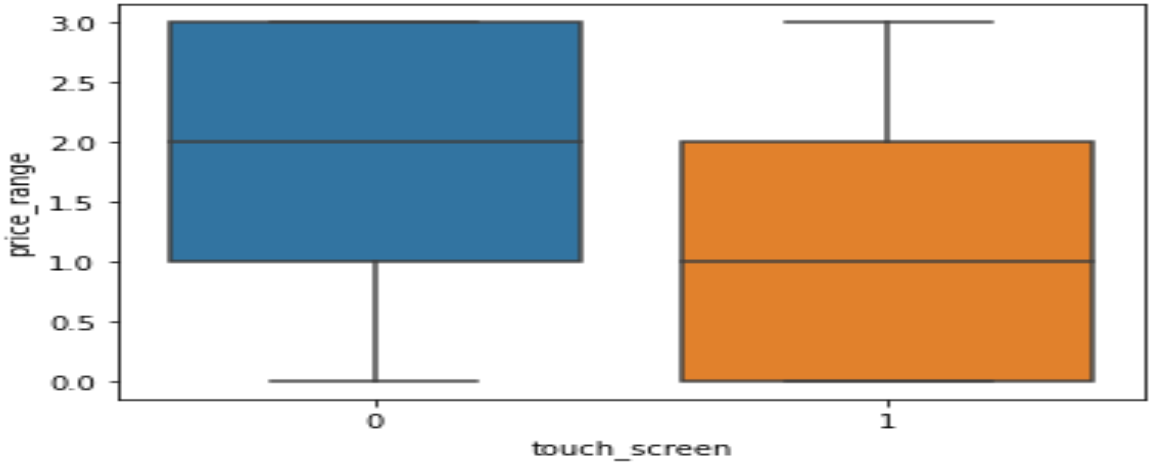

Width ranges mostly in 0-7

# Three G



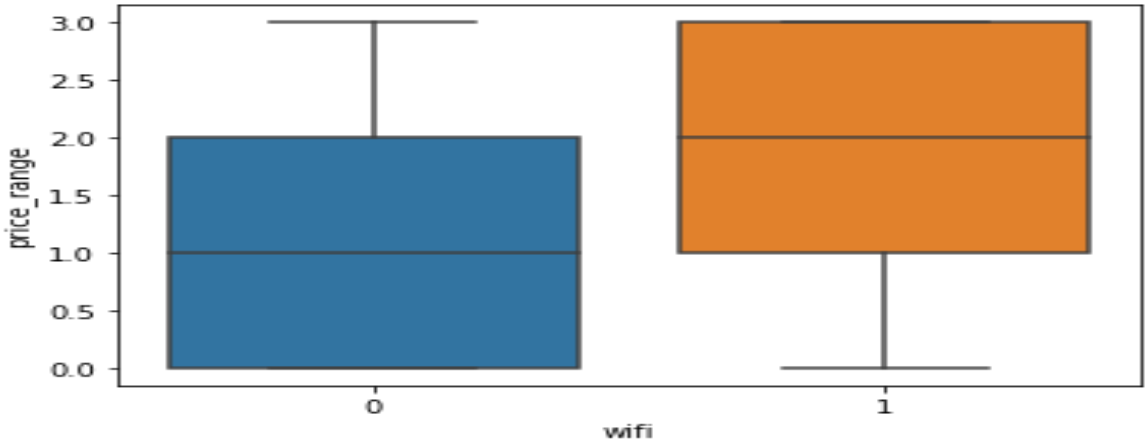

Price Range of 3G phones are considerably higher.

# Touch Screen




Price Range of touch screen phone is low.. Quite strange considering all the 4G,3G and Wifi phones are in higher price range

# Wifi




Price Range of wifi phones are considerably higher.

# Split Data

Before training our model on the dataset, we need to split the dataset into training and testing datasets.

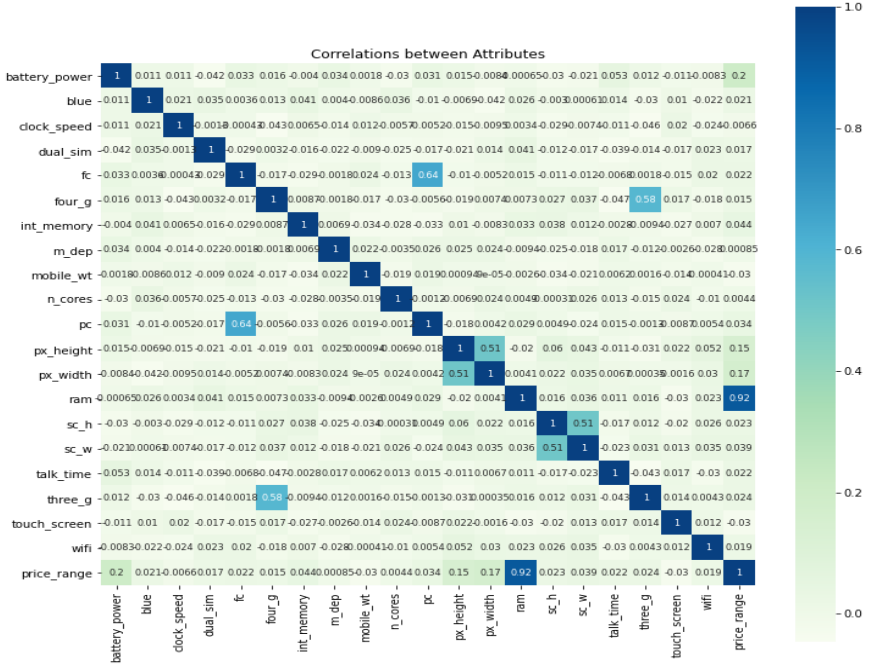
This is required to train our model on the major part of our dataset and test the accuracy of the model on the minor part.

We divide our dataset with a ratio of 80/20.

After splitting, there are 1600 data for training and 400 data for testing dataset.

# Correlation Analysis

- The most influential variable is RAM.
- Most of the variables have very little correlation to price range
- Primary camera mega pixels and front Camera mega pixels have correlation (it make sense because both of them reflect technology level of resolution of the related phone model) but they do not affect price range.
- Having 3G and 4G is somewhat Correlated.
- There is no highly correlated inputs in our dataset, so there is no multicollinearity problem.



# Logistic Regression

| | Precision | Recall | F1 score |
|---|---|---|---|
| 0 | 0.92 | 0.88 | 0.90 |
| 1 | 0.72 | 0.64 | 0.68 |
| 2 | 0.57 | 0.58 | 0.58 |
| 3 | 0.72 | 0.82 | 0.77 |

Accuracy = 0.73

# Decision Tree

| | Precision | Recall | F1 score |
|---|---|---|---|
| 0 | 0.92 | 0.89 | 0.90 |
| 1 | 0.79 | 0.74 | 0.76 |
| 2 | 0.72 | 0.80 | 0.76 |
| 3 | 0.90 | 0.88 | 0.89 |

Accuracy = 0.83

# Random Forest

| | Precision | Recall | F1 score |
|---|---|---|---|
| 0 | 0.97 | 0.91 | 0.94 |
| 1 | 0.85 | 0.91 | 0.88 |
| 2 | 0.88 | 0.85 | 0.86 |
| 3 | 0.92 | 0.94 | 0.93 |

Accuracy = 0.90

# K- Nearest Neighbors(KNN)

| | Precision | Recall | F1 score |
|---|---|---|---|
| 0 | 0.98 | 0.95 | 0.96 |
| 1 | 0.92 | 0.96 | 0.94 |
| 2 | 0.92 | 0.94 | 0.93 |
| 3 | 0.97 | 0.94 | 0.95 |

Accuracy = 0.95

# Best Model


**After training our dataset with four different model, we conclude that KNN is best model for our dataset. (via the highest accuracy score = 0.95). The best optimum K number is to be 9 for this dataset.**

# Best Hyperparameters

- Leaf_size = 30
- Metric = minkowski
- N_neighbors= 9
- Weights= uniform
- N_jobs = None
- Metric_params = None

# Conclusion

In this project we covered various aspects of the Machine learning development cycle. We observed that the data exploration and variable analysis is a very important aspect of the whole cycle and should be done for thorough understanding of the data.

Finally, we trained our model on optimal features with four different models namely logistic regression, Decision Tree, Random Forest and KNN.

We got KNN to be the best model with higher accuracy of 95%.

# Thank You