CAPSTONE PROJECT 3 (CLASSIFICATION)

SHARATH S

ABOUT THE DATASET

- The dataset contains information about different features of a mobile phone, and the price range.

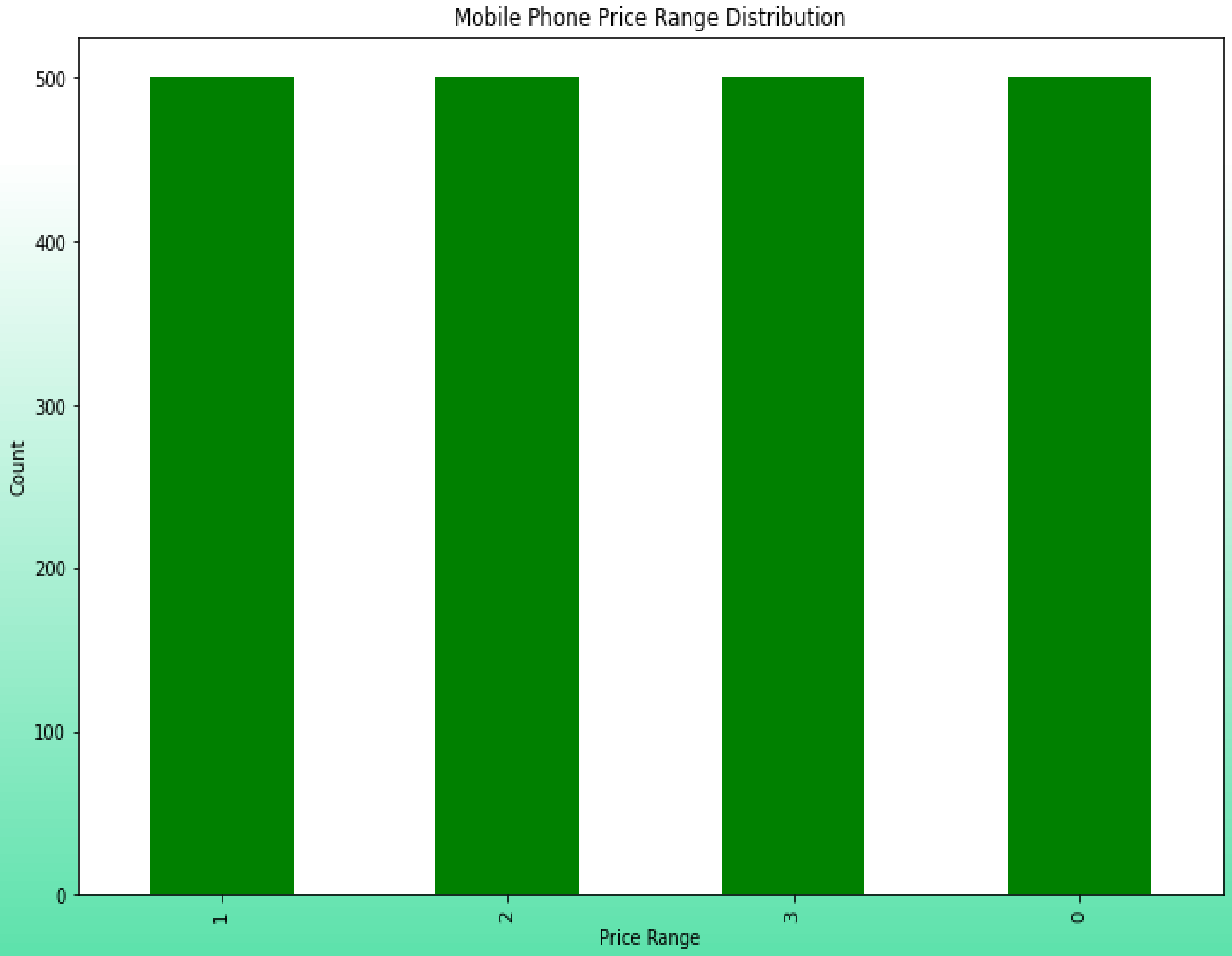
- The dependent variable is price range (0,1,2,3)

PROBLEM STATEMENT

- The mobile phone industry is highly competitive, and prices play a significant role in the purchasing decisions of customers. To remain competitive, companies needs to price their mobile phones effectively. However, determining the optimal price range for a mobile phone is challenging, given the numerous factors that influence pricing decisions
- The objective is to find out some relation between features of a mobile phone, and its selling price, and create a classification model to predict the price range.

EDA

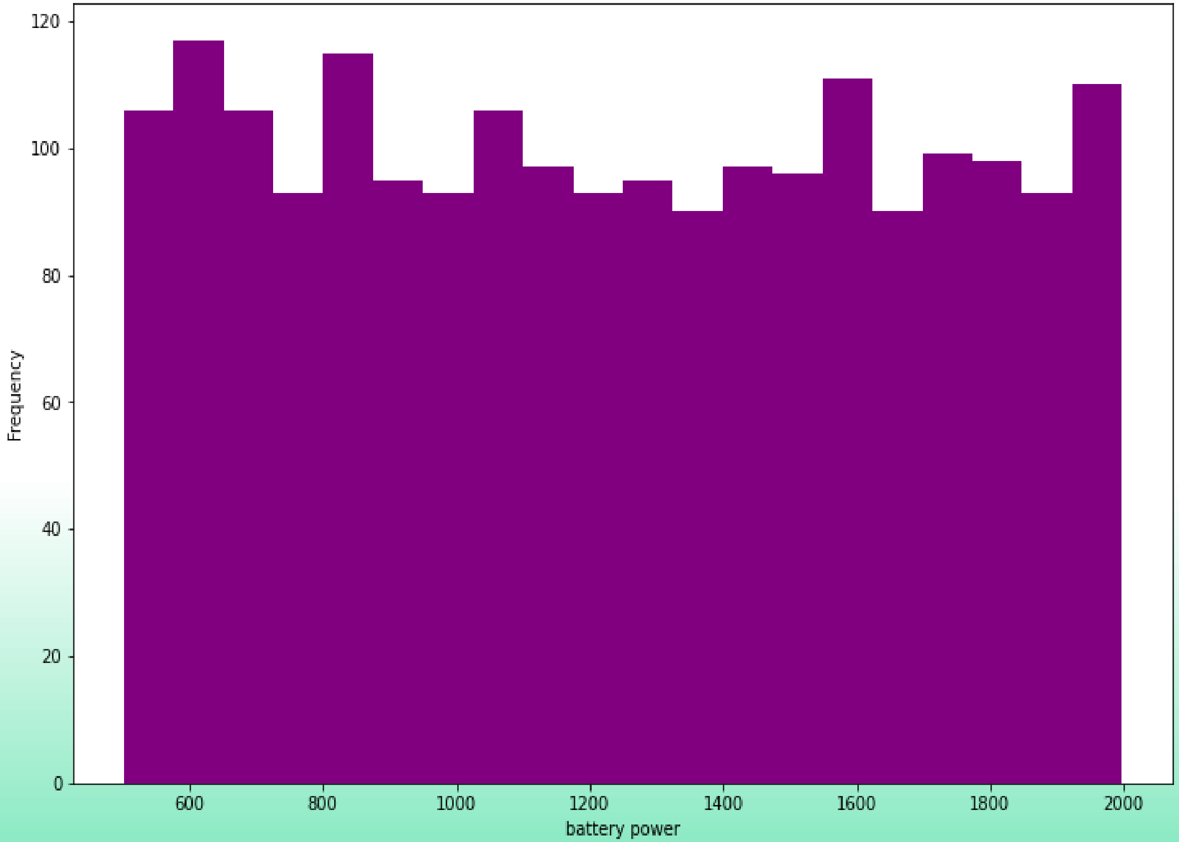
- Bar plot of price range



EDA

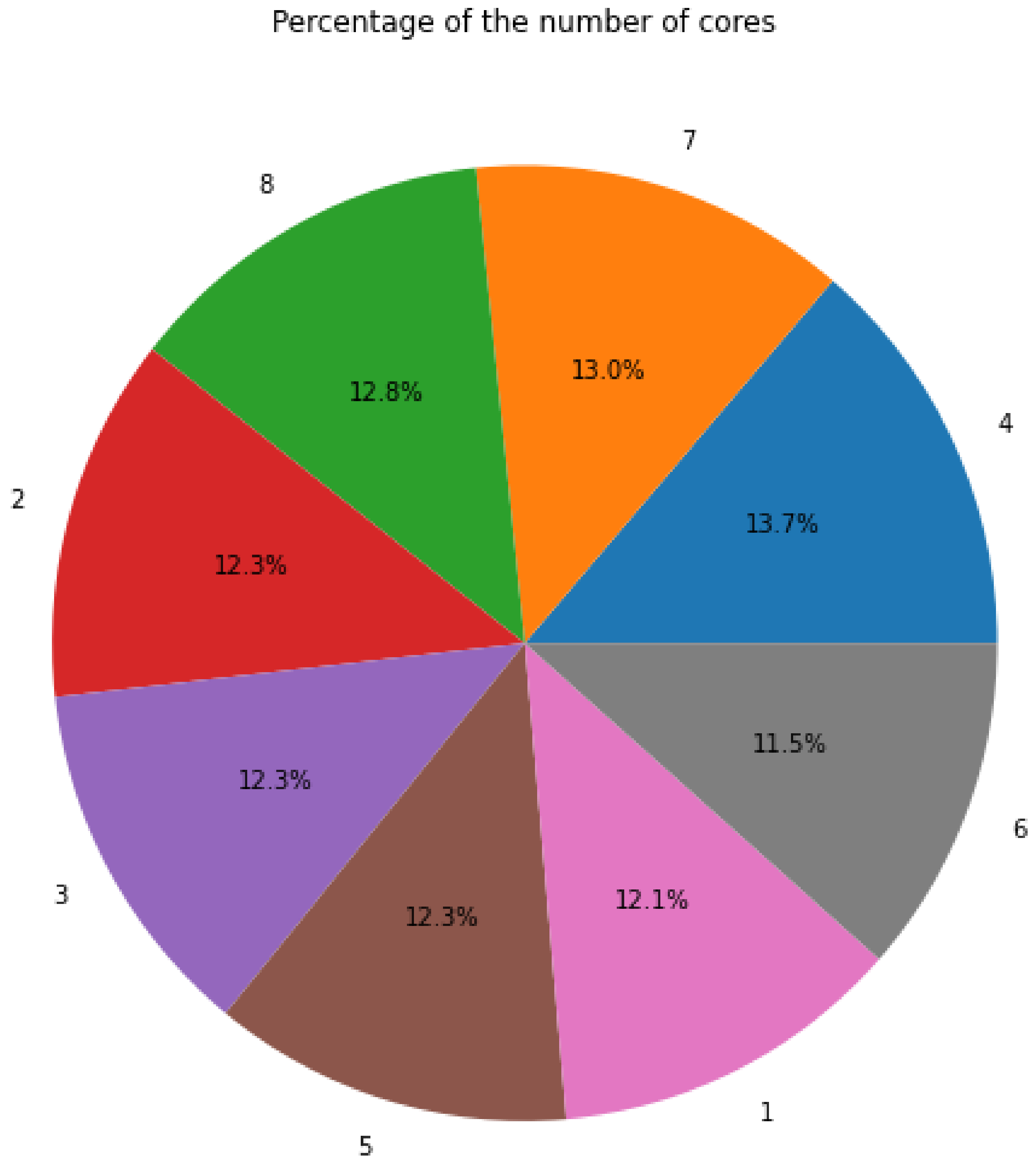
- Histogram of battery power

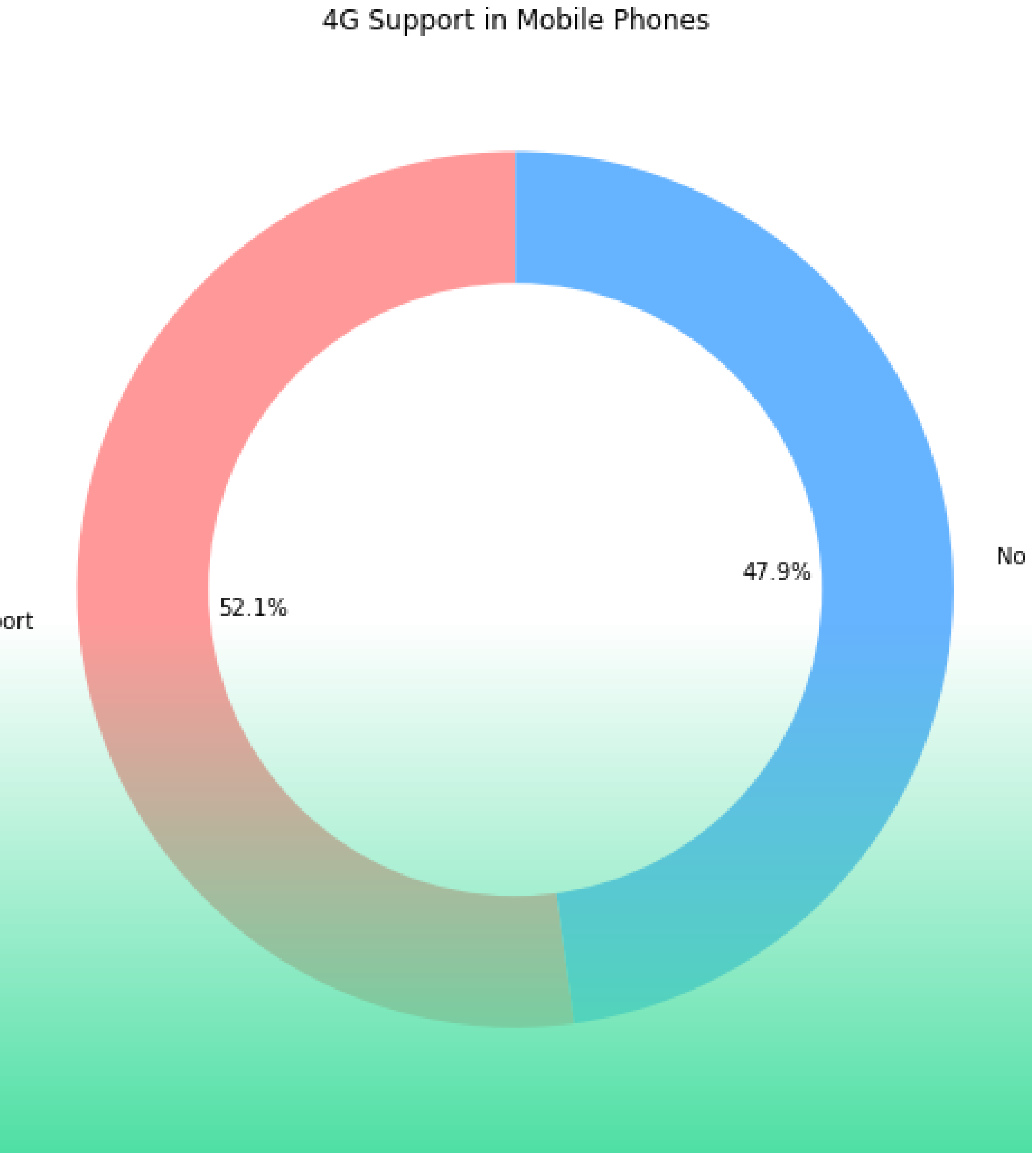
Phones with battery power of 600-650 have the most observations.



EDA

- Pie chart of percentage of number of cores.
- Phones with 4 cores have a higher coun t than phones with other cores and pho nes with 6 cores have the least count in the dataset.



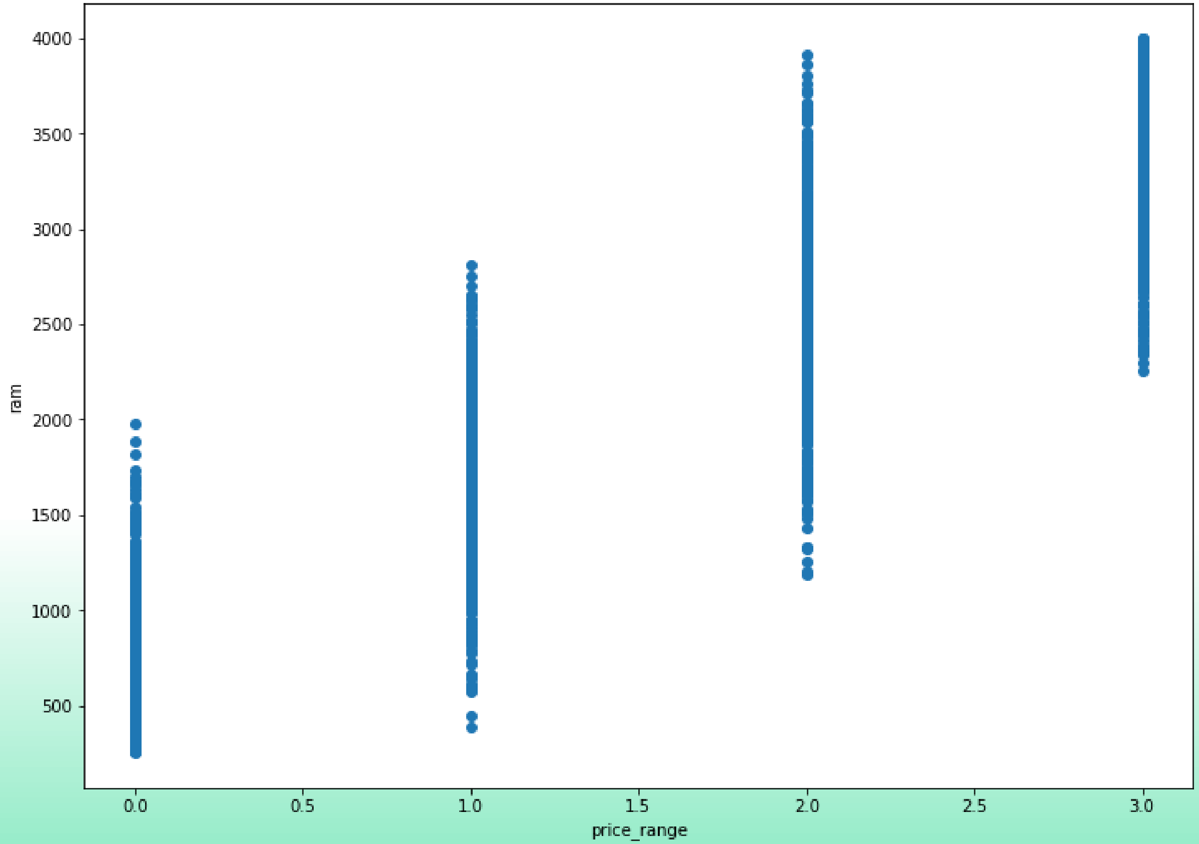


EDA

- Donut chart of 4G support.

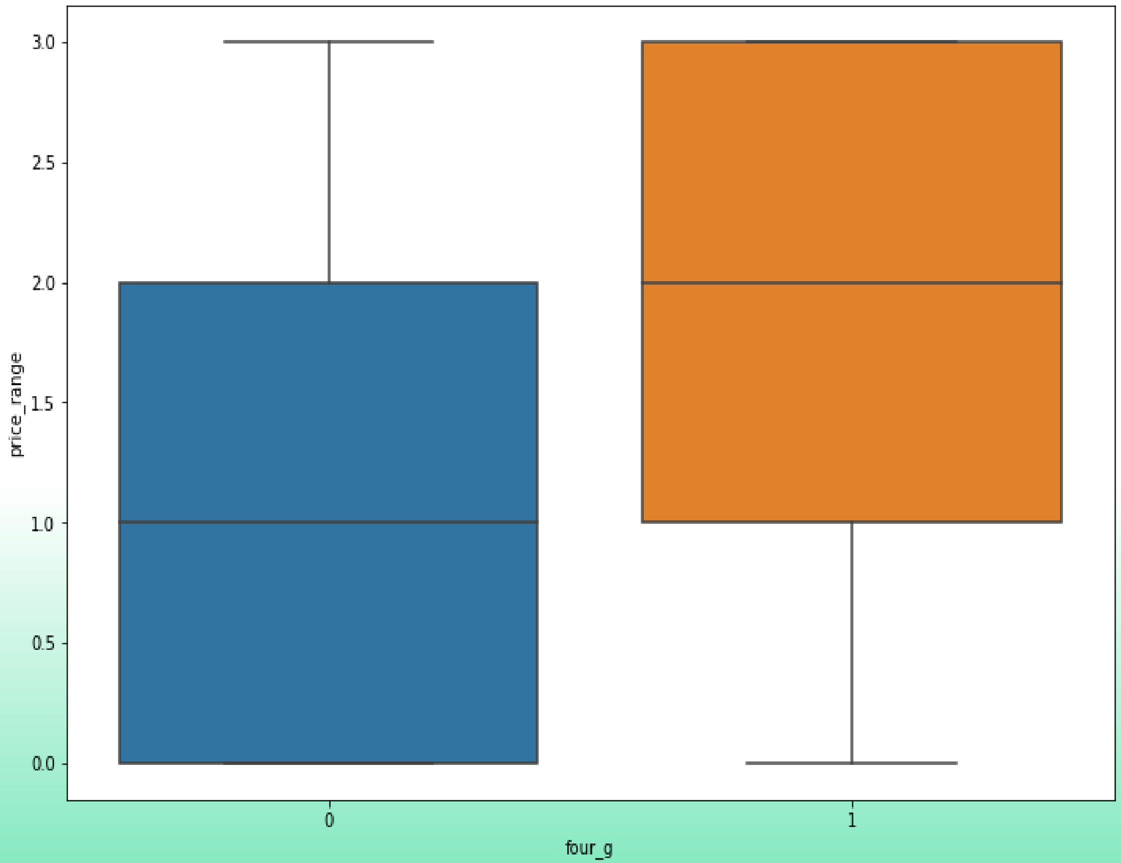
EDA

- Scatterplot of ram and price range



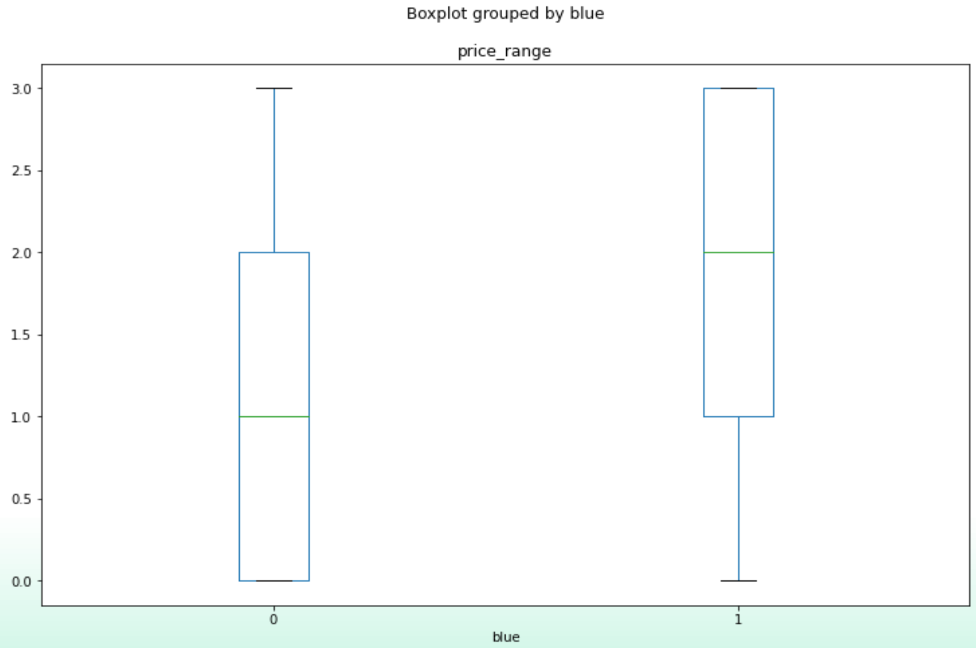
EDA

- Boxplot of Price range and 4G
- The median price of phones with 4G is higher than the median price of phones without 4G.



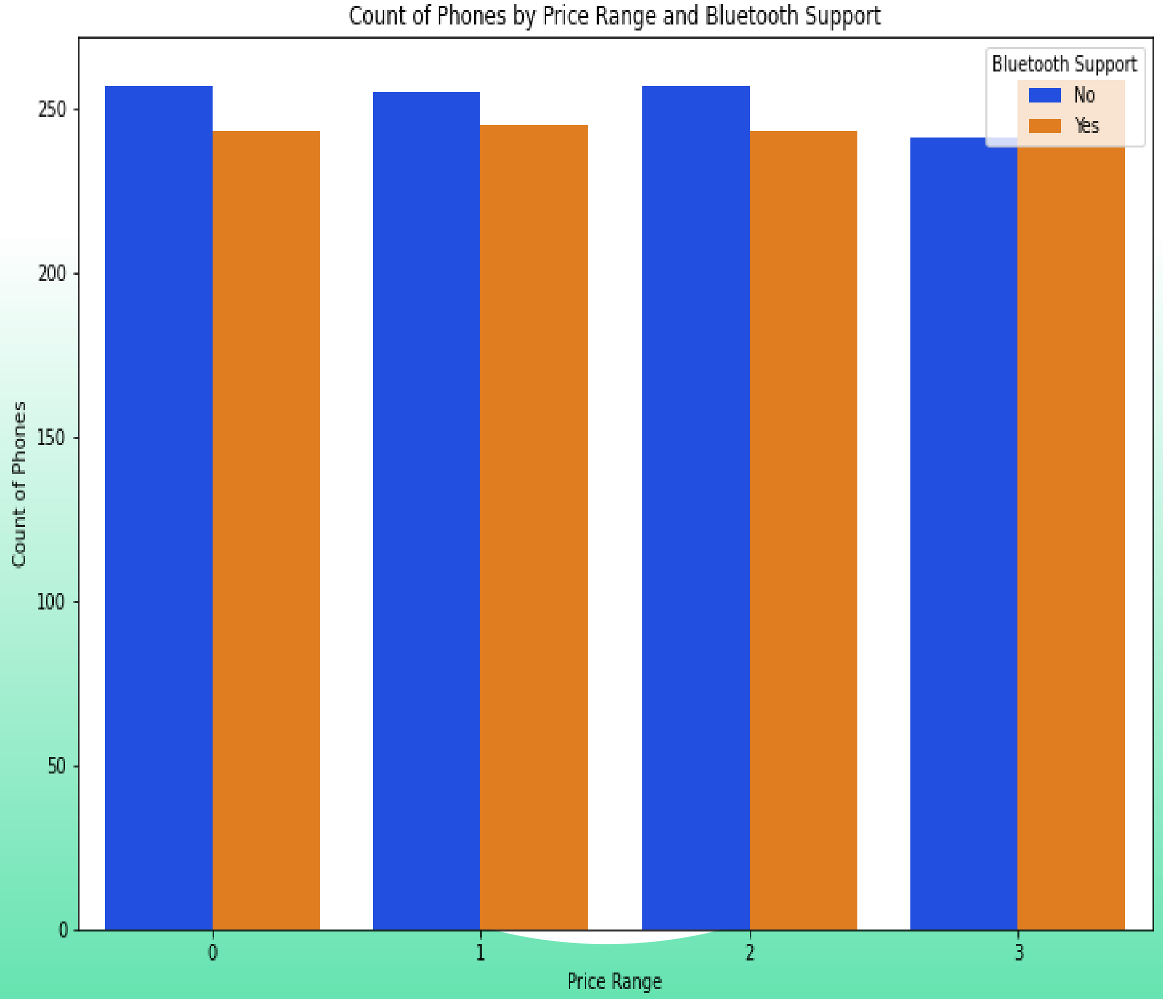
EDA

- Box plot of phones with Bluetooth and price range



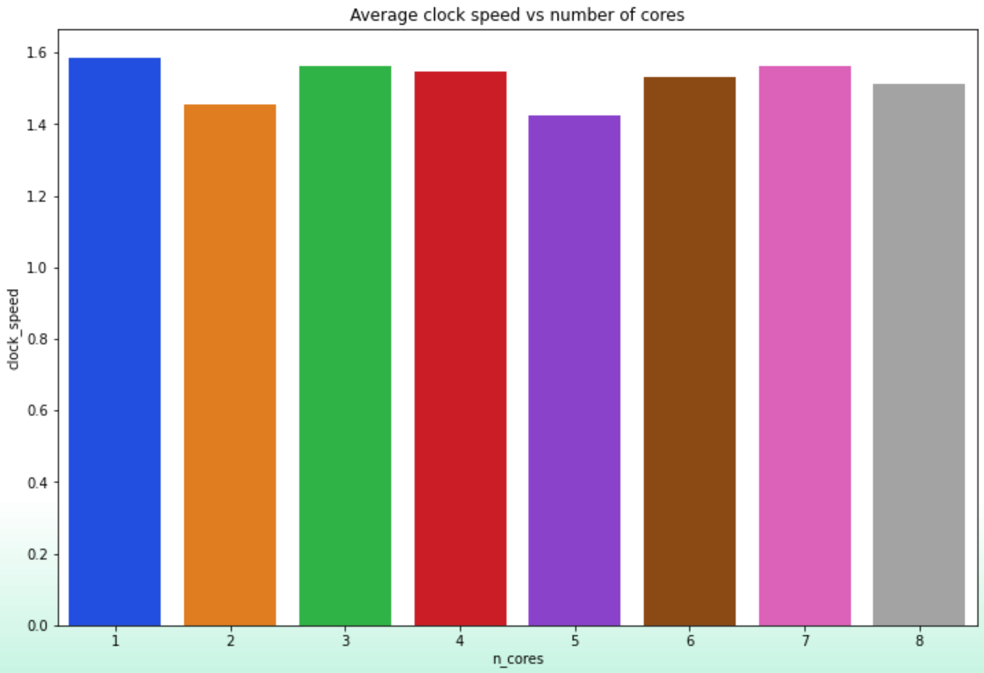
EDA

- Countplot of phones by price range and Bluetooth support
- Most of the phones which are at a higher price range have Bluetooth support.



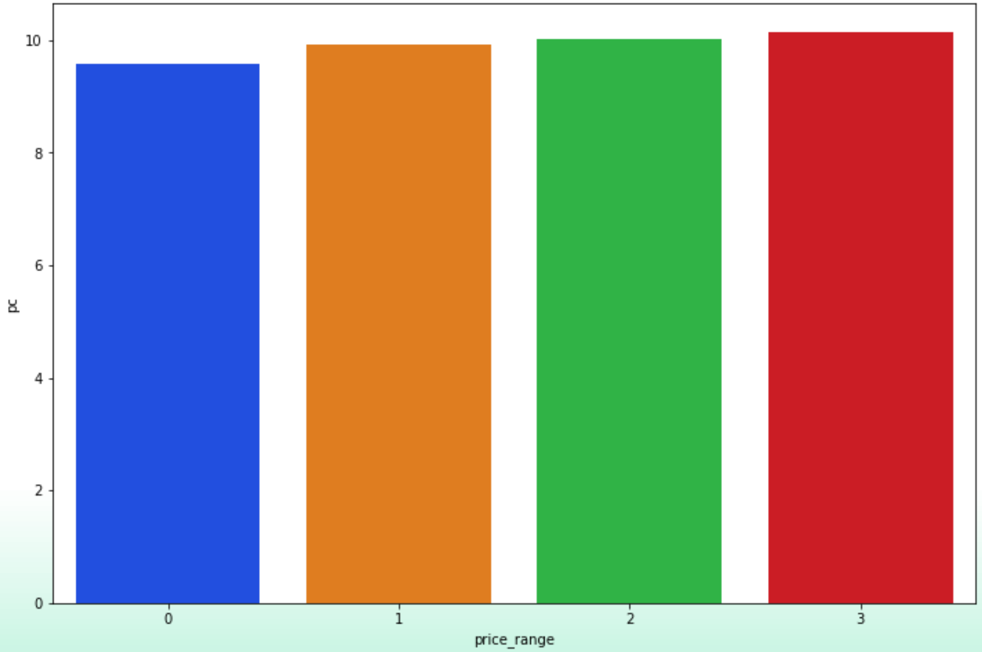
EDA

- Bar plot of number of cores vs avg clock speed.
- Phones with 1 core have the highest average clock speed and phones with 5 cores have the lowe st.



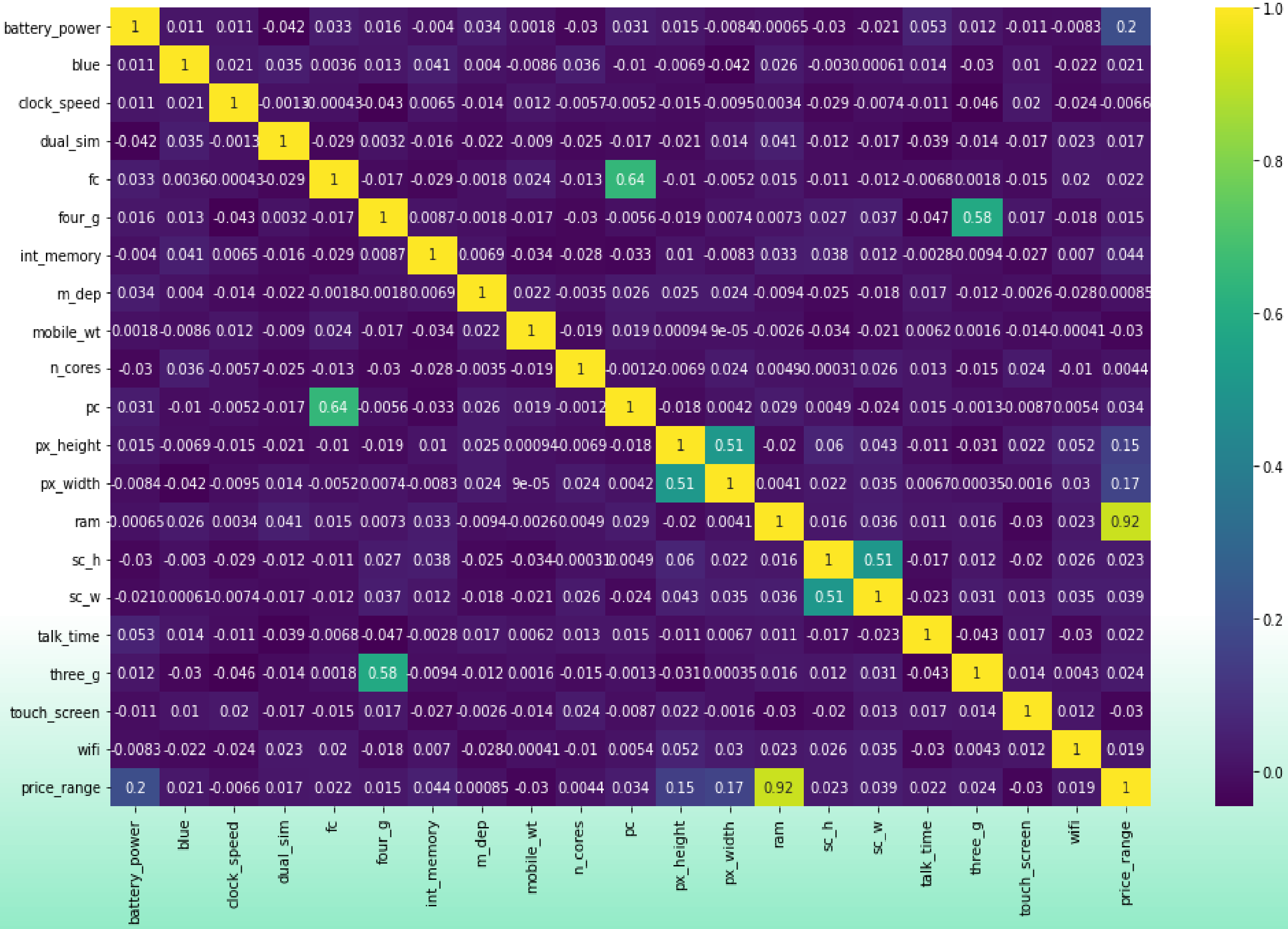
EDA

- Barplot of pc and price range
- From this chart, I can clearly see that costlier phones have cameras with higher megapixels.



E D A

- Correlation heatmap



FEATURE SELECTION USING VIF

Feature	VIF
battery_power	7.472227
blue	1.976321
clock_speed	4.066737
dual_sim	1.973756
fc	3.358165
four_g	3.186818
int_memory	3.837915
m_dep	3.786000
n_cores	4.451917

FEATURE SELECTION USING VIF

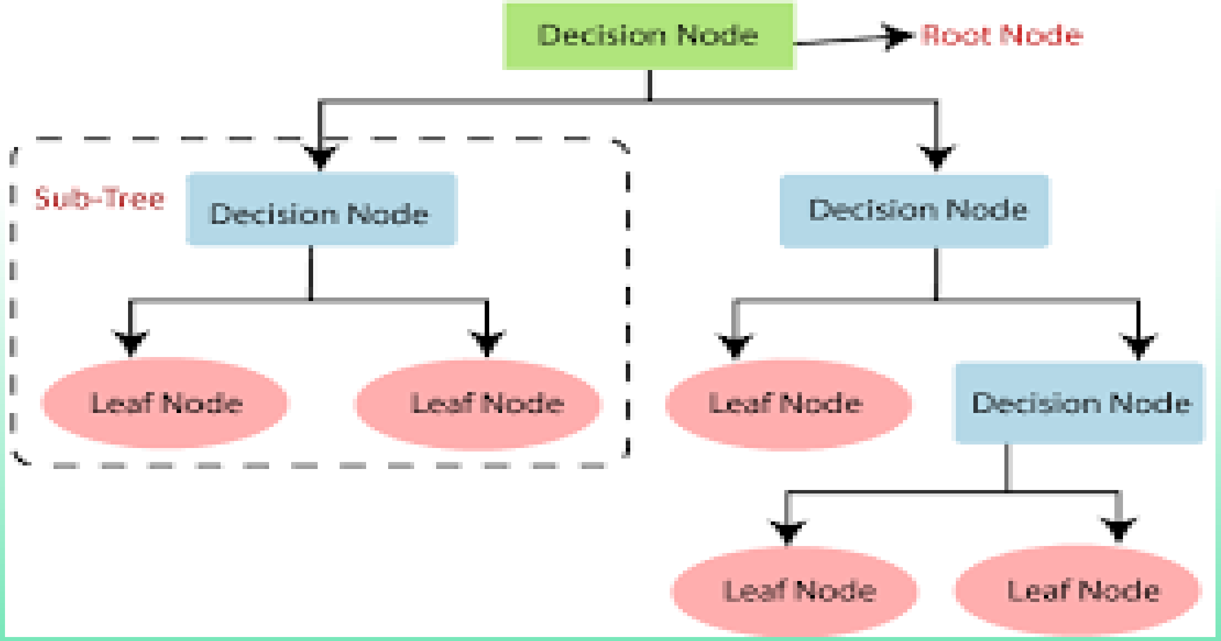
Feature	VIF
pc	5.840591
ram	4.555019
talk_time	4.629171
three_g	5.986670
touch_screen	1.978478
wifi	1.973050
px_area	2.155231
sc_area	2.083462

SCALING

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

MODEL 1: DECISION TREE

- The ML model I used is Decision Tree. It is a nonparametric supervised learning algorithm. With a hier archical, tree like structure.

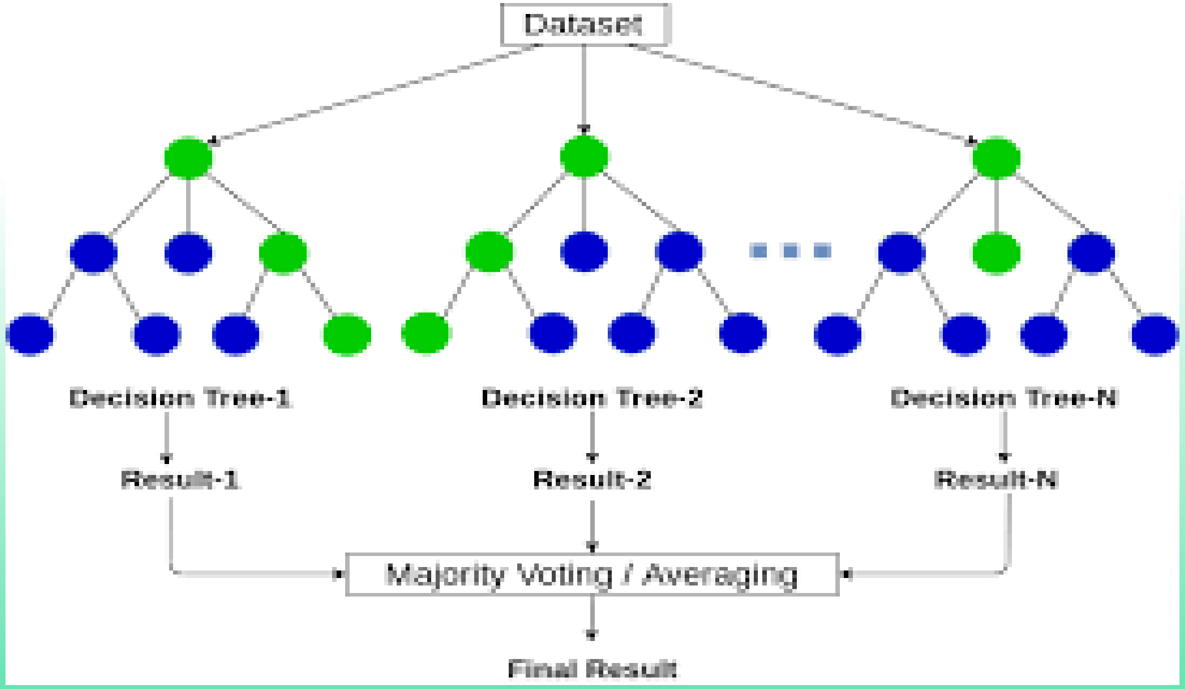


EVALUATION METRICS FOR DECISION TREE

	precision	recall	f1-score	support
0	0.88	0.92	0.90	90
1	0.83	0.81	0.82	102
2	0.78	0.85	0.81	97
3	0.95	0.85	0.90	103
accuracy			0.86	392
macro avg	0.86	0.86	0.86	392
weighted avg	0.86	0.86	0.86	392

MODEL 2: RANDOM FOREST

- It is an ensemble learning method which builds multiple decision trees on different random subsets of the training data and features.
-

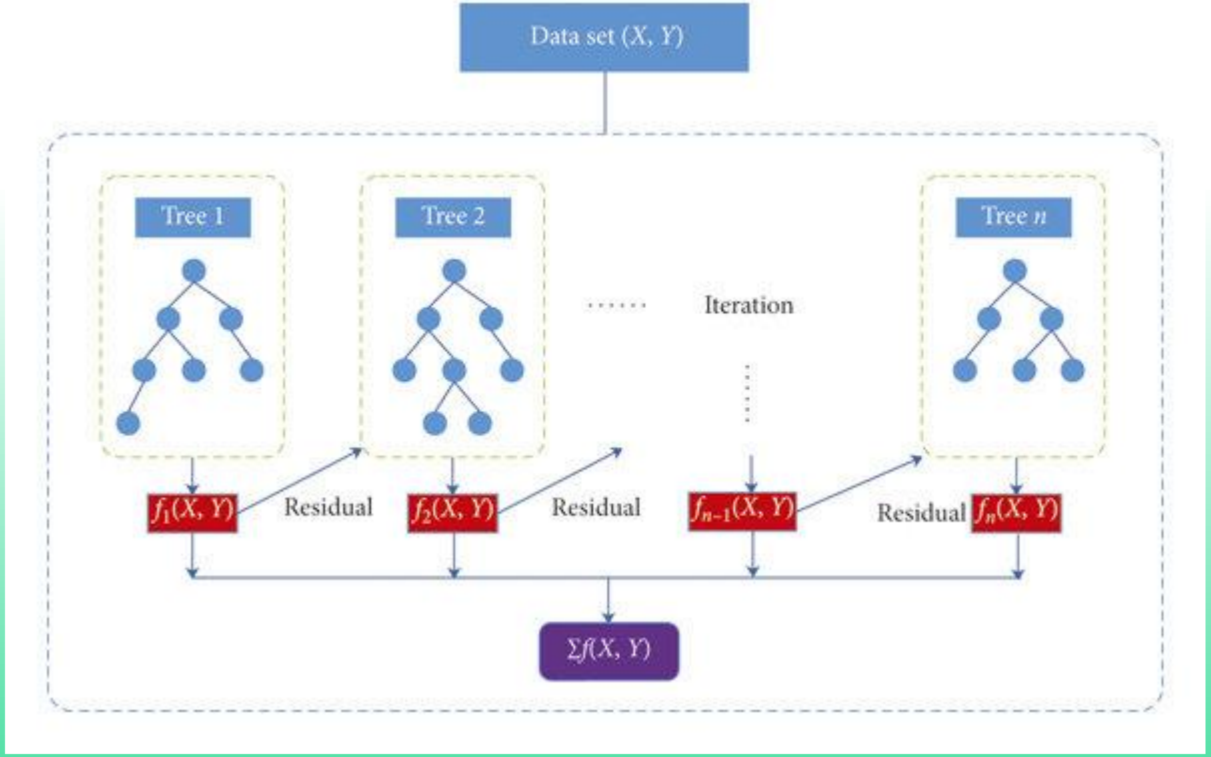


EVALUATION METRICS FOR RANDOM FOREST

	precision	recall	f1-score	support
0	0.85	0.92	0.88	90
1	0.85	0.80	0.83	102
2	0.87	0.91	0.89	97
3	0.98	0.92	0.95	103
accuracy			0.89	392
macro avg	0.89	0.89	0.89	392
weighted avg	0.89	0.89	0.89	392

MODEL 3: XGBOOST

- The algorithm works by iteratively training a sequence of decision trees, where each subsequent tree i s trained to correct the errors of the previous one



EVALUATION METRICS FOR XGBOOST

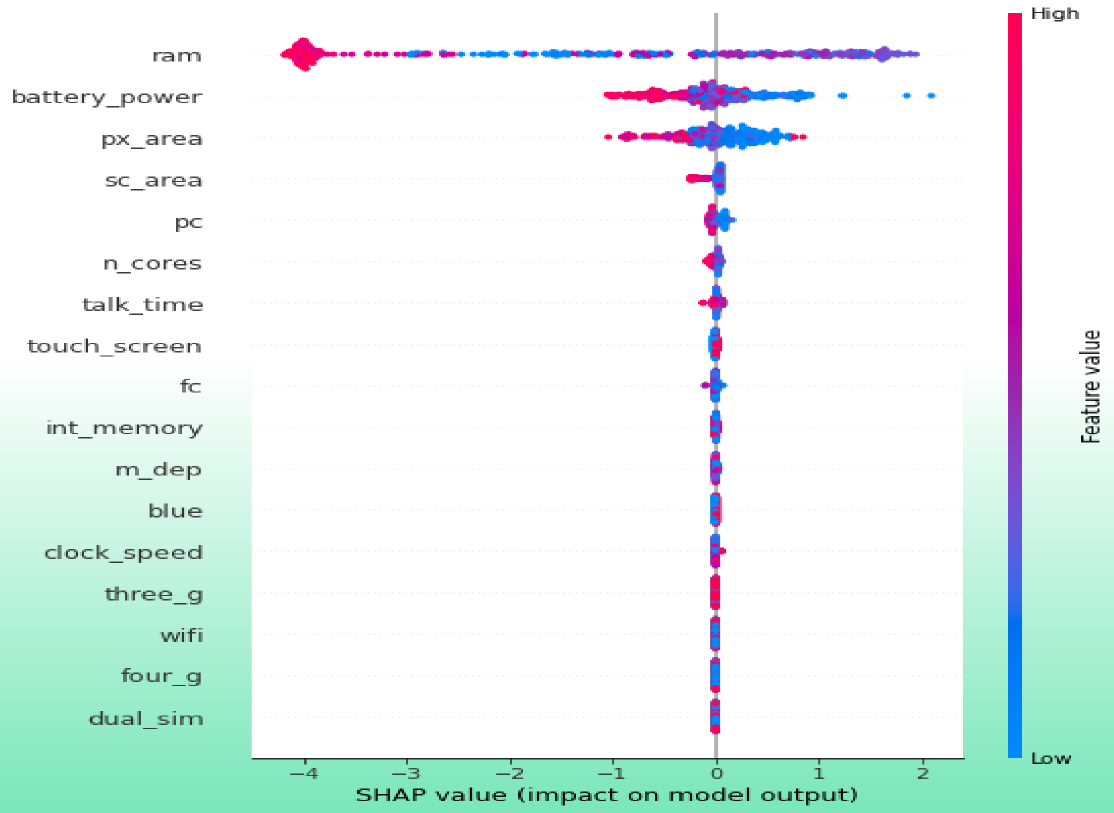
	precision	recall	f1-score	support
0	0.92	0.94	0.93	90
1	0.88	0.89	0.89	102
2	0.88	0.89	0.88	97
3	0.96	0.92	0.94	103
accuracy			0.91	392
macro avg	0.91	0.91	0.91	392
weighted avg	0.91	0.91	0.91	392

MODEL CHOSEN

- I would choose the XGBoost model as my final prediction model since it give me the best score out of all the other models, in terms of the evaluation metrics that I chose.

- I would choose accuracy as the most important metric for a positive business impact. The reason for this is that in this scenario, the cost of misclassification of a phone's price range is generally equal for all price ranges, and the classes are not imbalanced. Accuracy score should be used over precision and recall when the classes in the dataset are balanced

MODEL EXPLAINABILITY USING SUMMARY PLOT



THANK YOU