Capstone Project-3

Mobile Price Range Prediction

Supervised Machine Learning (Classification)

BY

Girish R

Contents

1. Problem Statement
2. Knowing the Dataset
3. Data Cleaning
4. EDA (Exploratory Data Analysis)
5. Data Splitting
6. ML Models Implementations
7. Conclusion

Problem Statement

The problem statement is to predict the price range of mobile phones based on the features available (price range indicating how high the price is).Here is the description of target classes :

- 0 - Low cost Phones
- 1 - Medium cost phones
- 2 - High cost phones
- 3 – Very High cost phones

This will basically help companies to estimate price of mobiles to give tough competition to other mobile manufacturer.
Also, it will be useful for consumers to verify that they are paying Best price for a mobile.

Knowing the Dataset:

Total Rows= 2000
Total features=21

Independent Variables

- **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 and Px_width** - Pixel Resolution Height and width.

- **Ram -** Random Access Memory in Mega Bytes.
- **Sc_h and Sc_w -** Screen Height and 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) and3(very high cost).

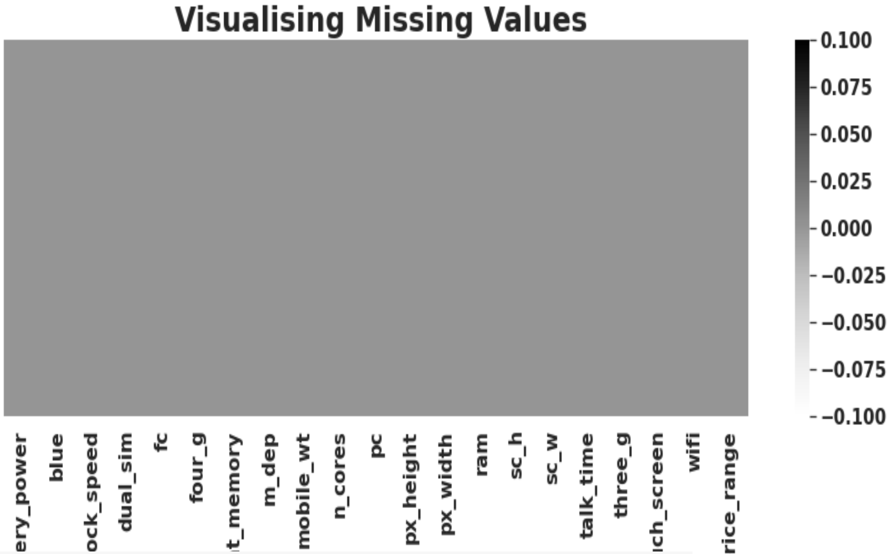
Dependent variables.

Price_range :This is the target variable with value of

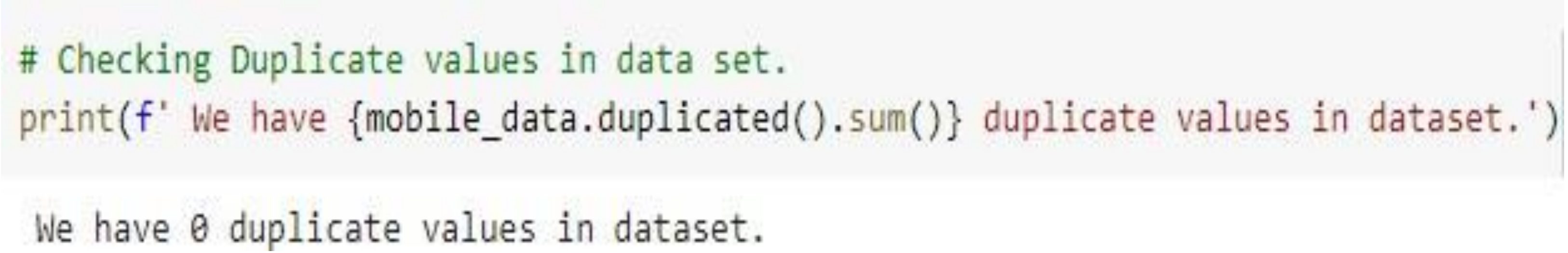
- 0 (low cost) ,
- 1 (medium cost) ,
- 2 (high cost) ,
- 3 (very high cost) .

Data Cleaning:

We analysed the dataset for duplicate values, null values, and missing values and noticed that none were present, implying that there were no duplicate, missing, or null values in the given dataset.



- **Zero Missing values after handling mismatch from the data.**
- **0 duplicates.**



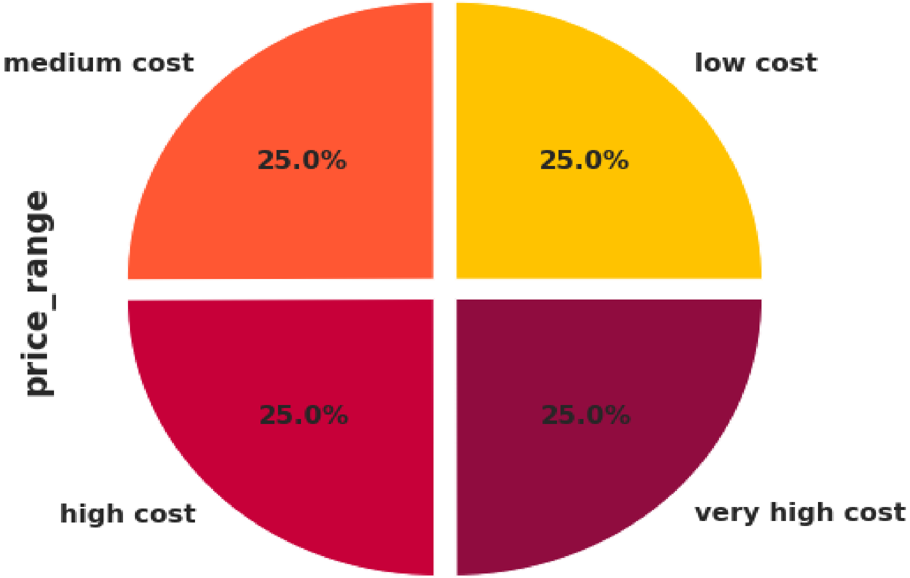
```
# Checking Duplicate values in data set.
print(f' We have {mobile_data.duplicated().sum()} duplicate values in dataset.')
```

```
We have 0 duplicate values in dataset.
```

EDA (Exploratory Data Analysis):

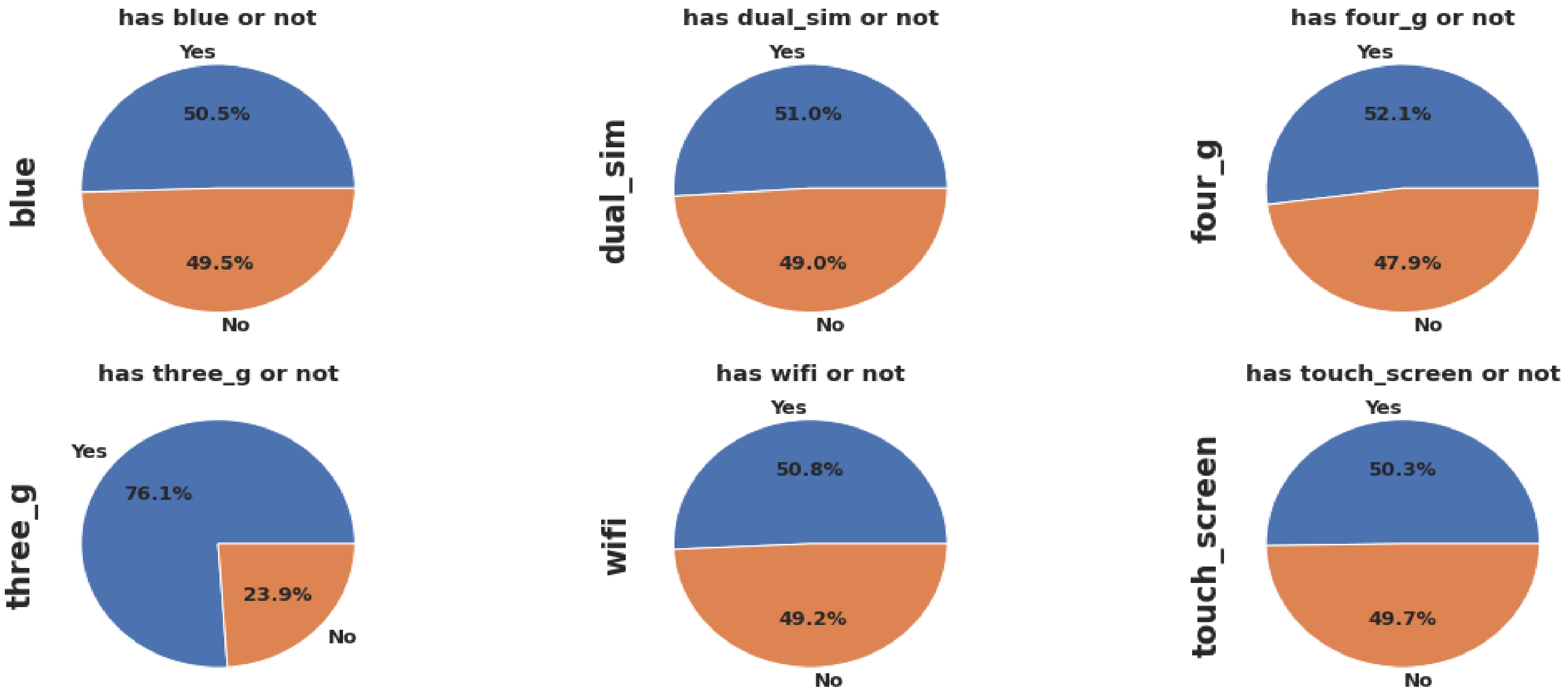
According to the pie chart,

- ✓ There are about 500 mobile data that are low cost, which is 25%, 500 mobile data that are medium cost, which is 25%, 500 mobile data that are high cost, which is 25%, and 500 mobile data that are very high cost, which is 25% of the total mobile data given in the dataset.
- ✓ In other words, all mobile price range categories are equally important, and we must focus on all price range people to maximize our sales and profit.

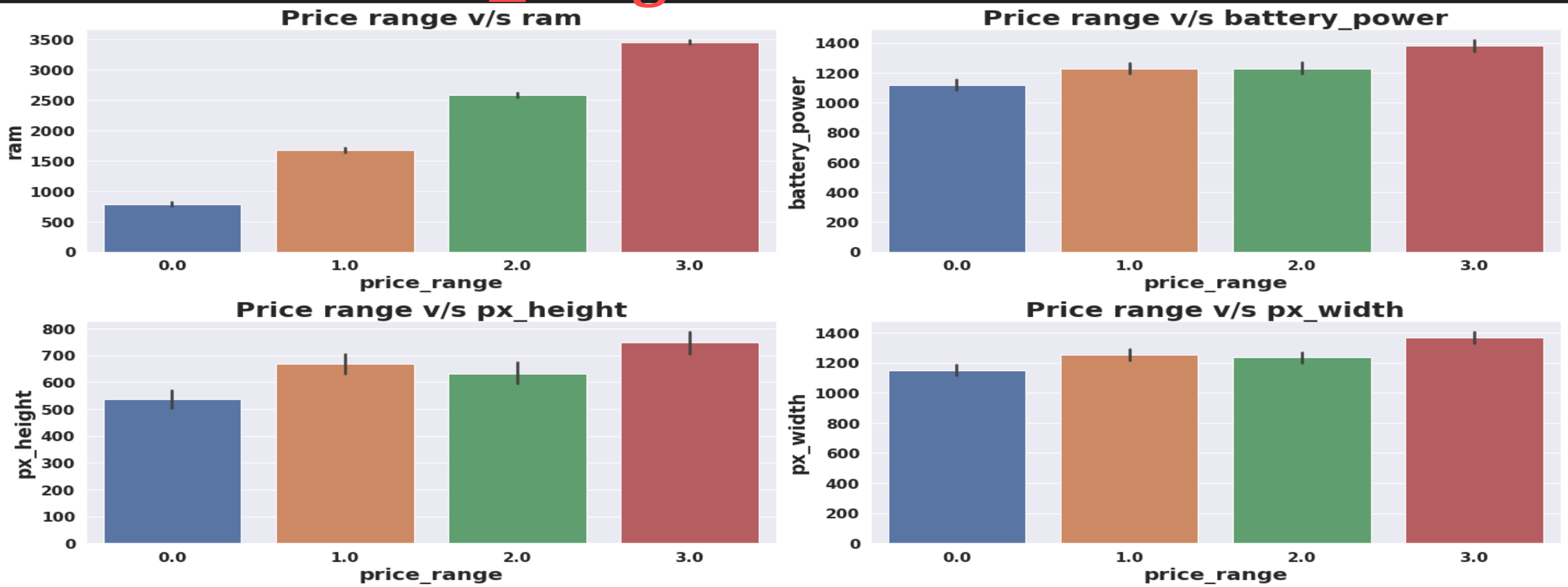


Univariate Analysis of Categorical columns.

- Percentage Distribution of Mobiles having Bluetooth ,dual sim, 4G, wifi and touchscreen are almost 50 %.
- Very few mobiles(23.8%) do not have Three_g.



Price_range vs Other Features



Price range v/s ram

Price range v/s battery_power

Price range v/s px_height

Price range v/s px_width

➢Mobiles having RAM more than 3000MB falls under Very high cost category. As RAM increases price range also increases.

➢Mobiles having RAM less than 1000 MB falls under low cost category.

➢Mobiles with battery power more than 1300 mAh has very high cost. And Mobiles with battery power between 1200 and 1300 mAh falls under medium and high cost category.

➢Mobiles with more than 700 pixel height and width more than 1300 has very high cost.

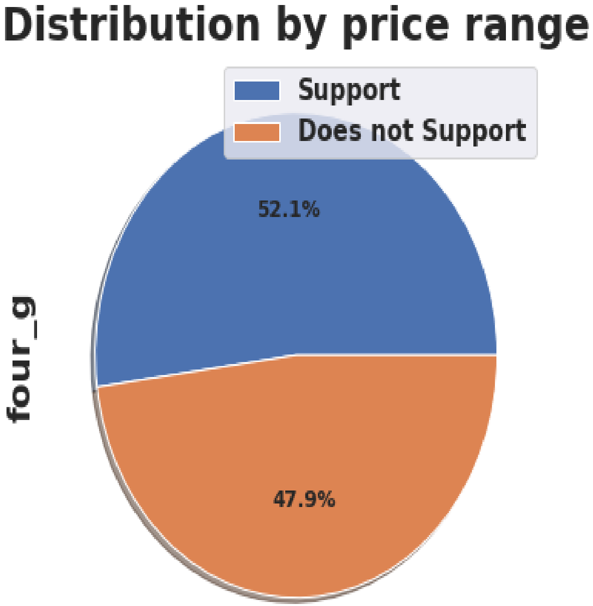
Distribution by Price range:

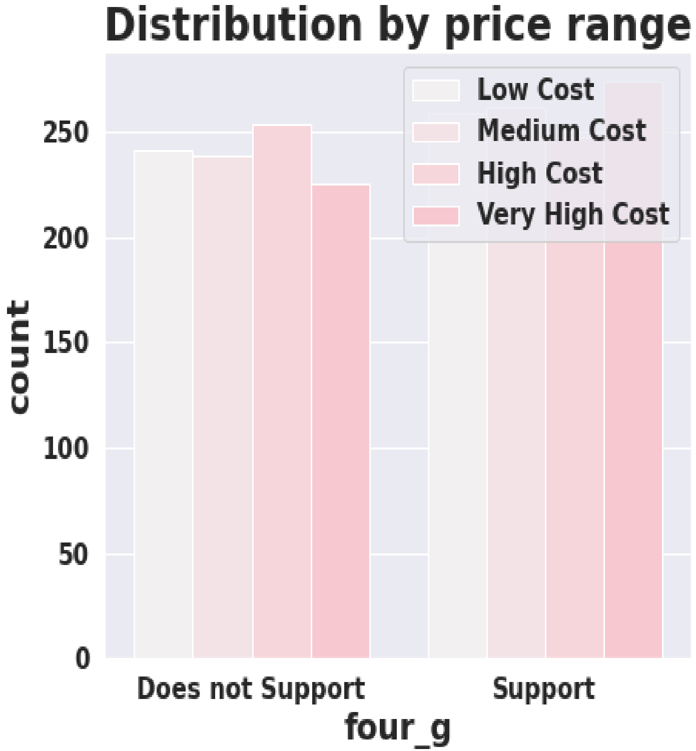
- The pie chart shows the percentage of mobile phones that support or do not support a specific binary feature.
- The bar chart shows the count of mobile phones for each binary feature based on the price range

❑ four_g

1.support=52.1%

2.does not support=47.9%

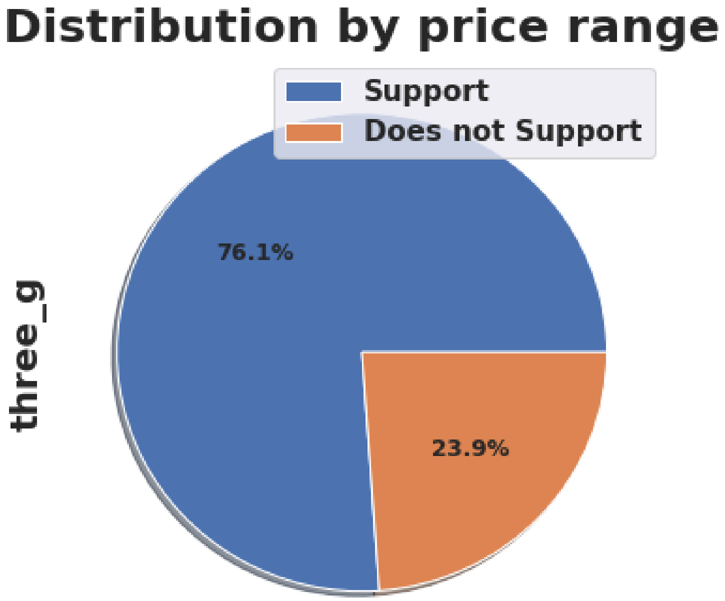


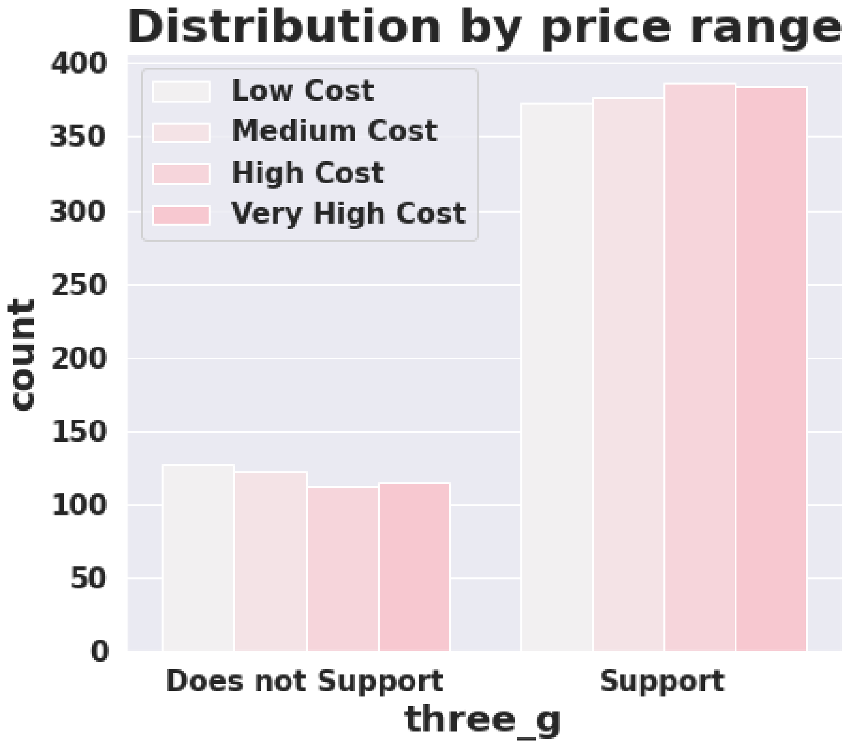


❑ three_g

1.support=76.2%

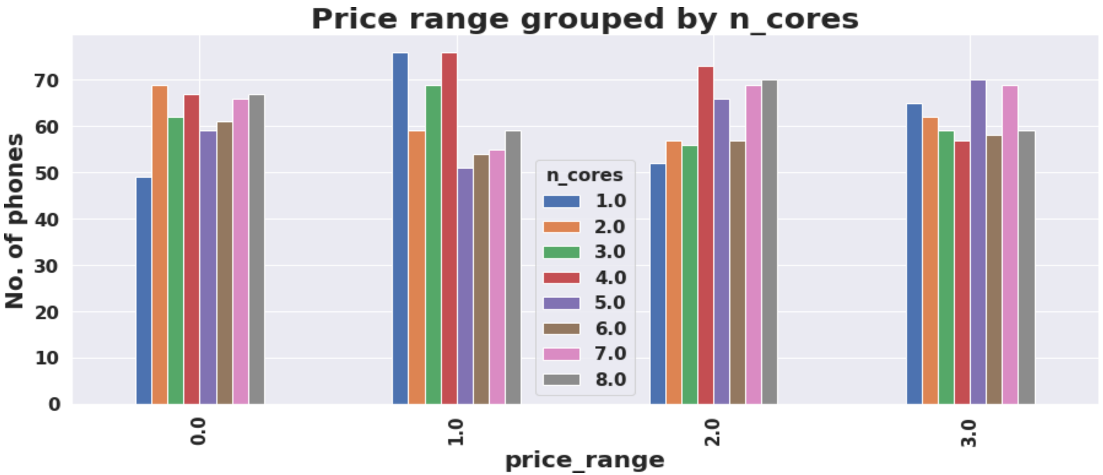
2.does not support=23.8

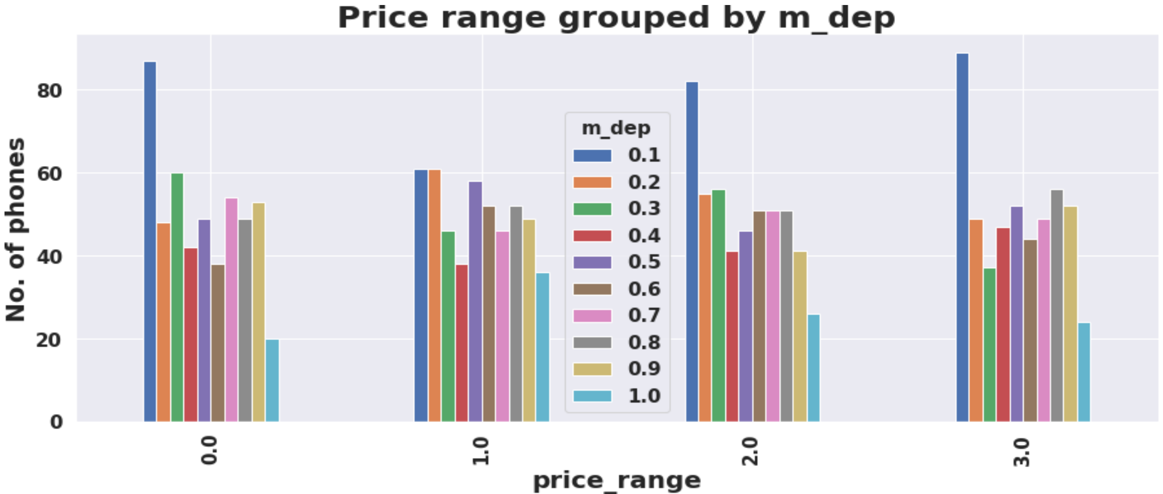




Bivariate and Multivariate Analysis:

- There are very few mobiles in price range 0 and 1 with lesser no of cores.
- Most of the mobiles in price range 2 and 3 are with high no of cores.
- Number of phones with less thickness is high and count of phones with high thickness is low.

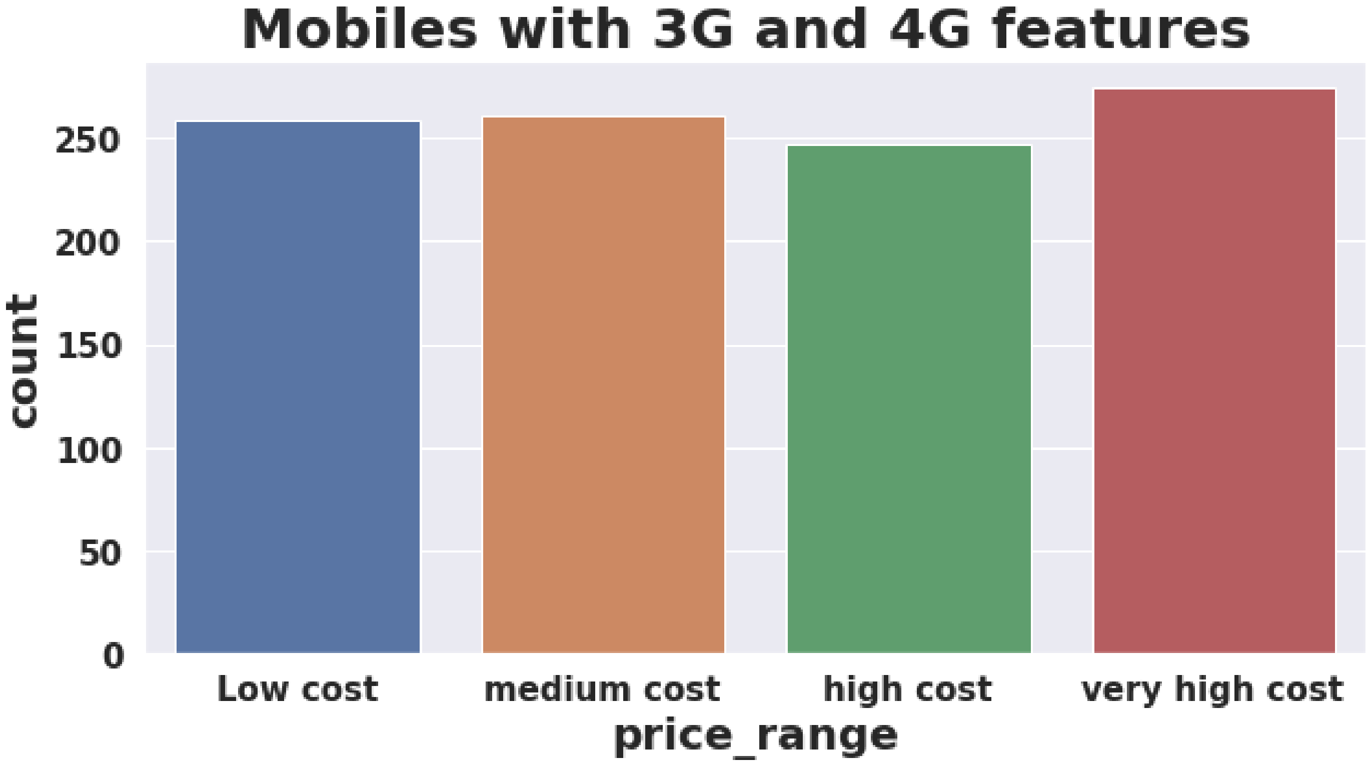




Mobiles with 3G and 4G Features

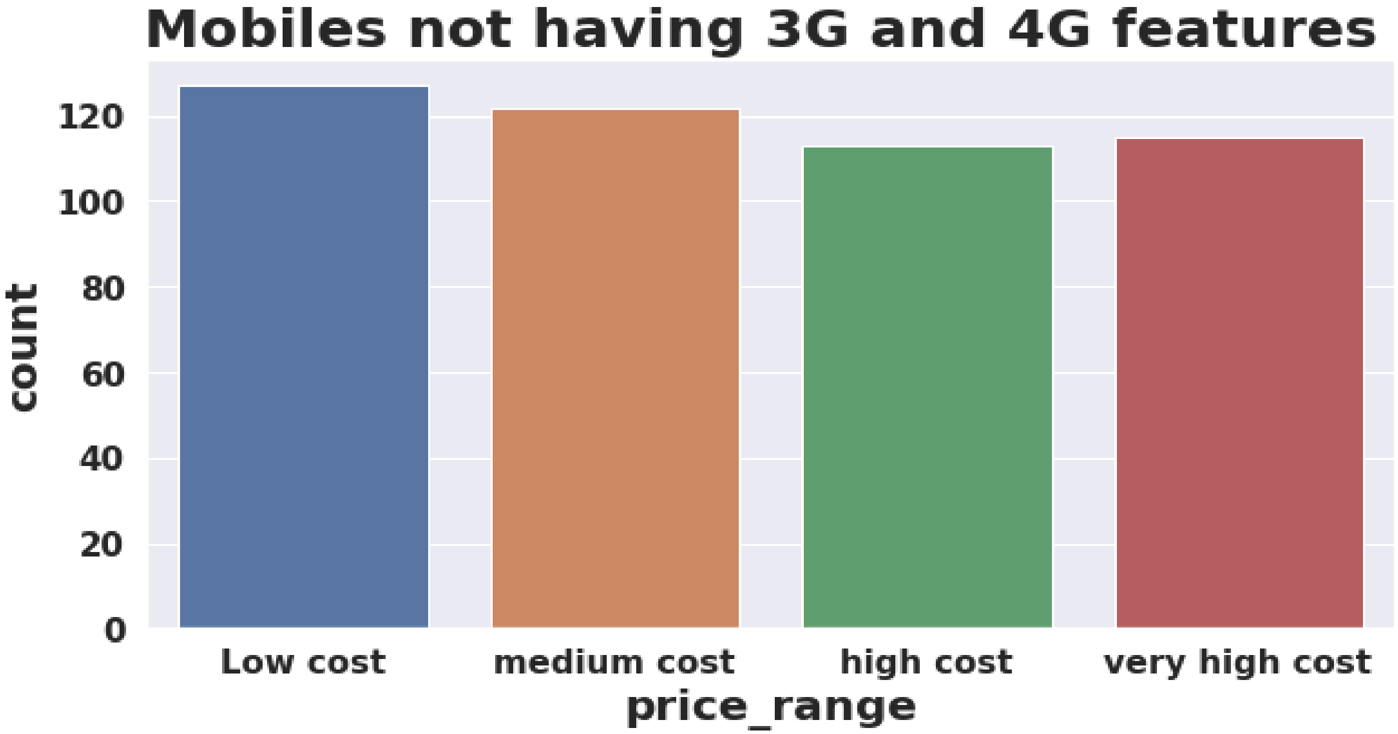
From the below graph,

As we can see from low cost to very high cost mobiles which have both 4G and 3G feature in it



Mobiles with 3G and 4G Features

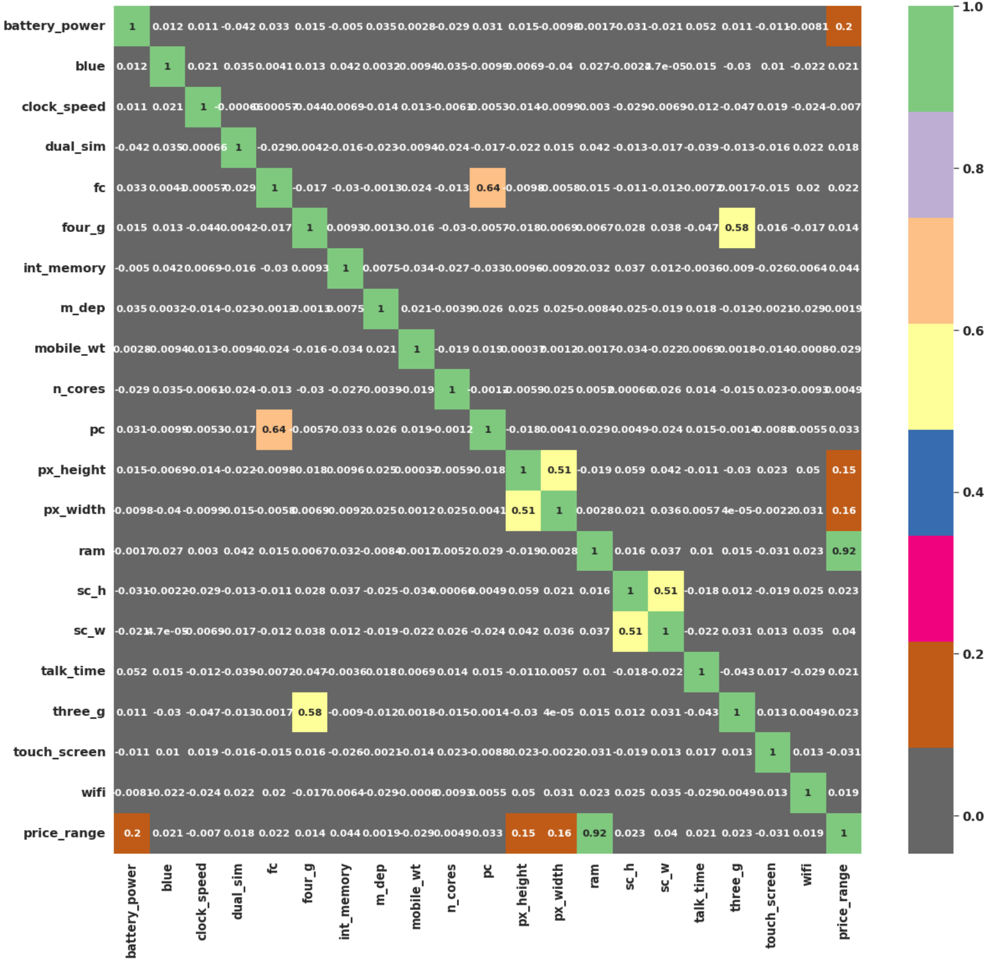
- It goes without saying that low-cost mobile phones will lack 3G and 4G capabilities.
- 5G may be available in high-end smartphones. As we all know, technology is constantly evolving.



Correlation Heatmap visualization code

From the correlation heatmap,

- ✓ RAM has strong positive correlation with the Price range and we know that Mobiles with high RAM are very costly. Thus RAM increases price range also increase.
- ✓ Battery power also has positive correlation with the price range. Generally mobiles having high prices comes with good battery power.
- ✓ Also px_heightand px_width(Pixel Resolution Height and width) are positively correlated. Generally High price range mobiles have good resolutions.
- ✓ Four_gand Three_gare highly positvelycorrelated. Nowdaysmost of the smart mobiles has both type of options. This could be the reason that they are correlated.
- ✓ primary camera i.epc and front camera fcare positively correlated.
- ✓ sc_hand sc_ware positively correlated.



Data Splitting :

Before building models we performed the train test split. We kept 25% of the data for testing and remaining 75% of the data for training the model.

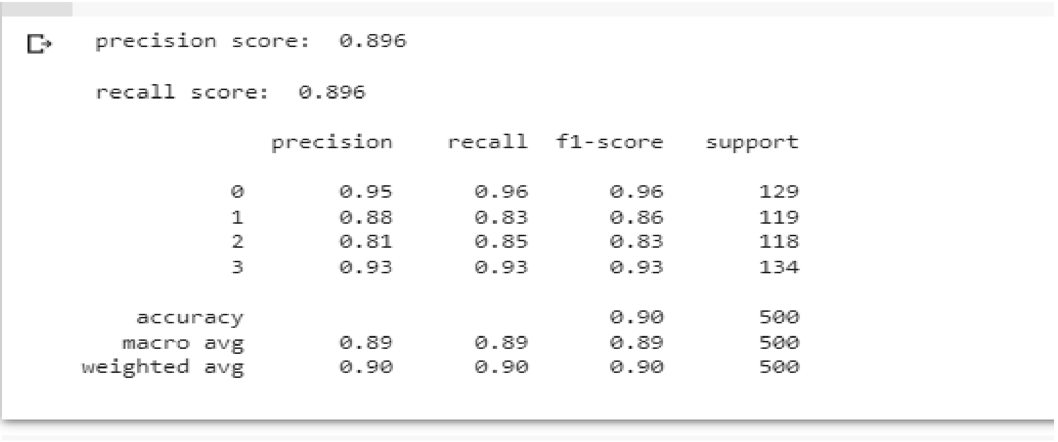
We compared 7 algorithms and evaluated them based on the overall accuracy score and the recall of the individual classes.

- Accuracy is the ratio of the total number of correct predictions and the total number of predictions.
- The recall is the measure of our model correctly identifying True Positives.

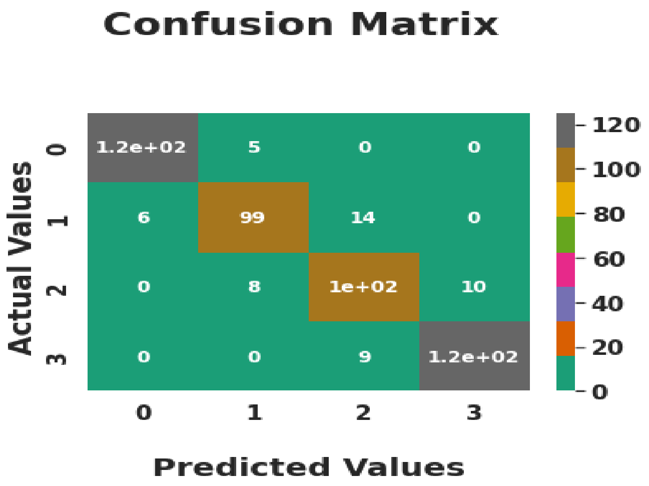
ML Model Implementation

1) **Logistic Regression**
2) **K-nearest Neighbour classifier**
3) **Naïve Bayes Theorem**
4) **Support Vector Machine(SVM)**
5) **Decision Tree**
6) **Random Forest classifier**
7) **XG Boost Classifier**

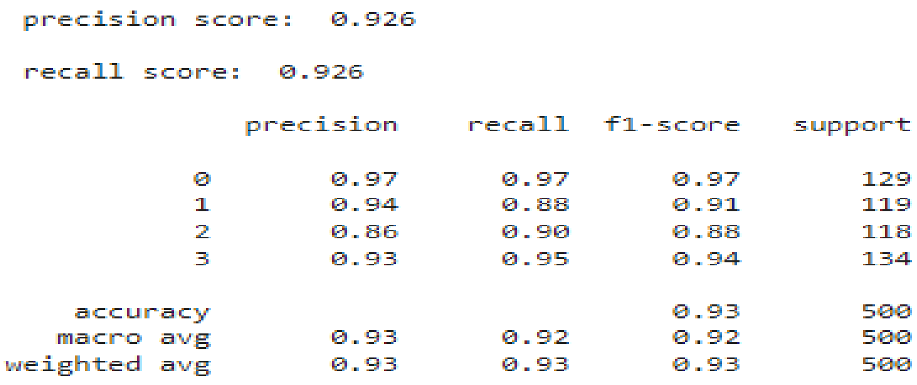
Logistic Regresssion:



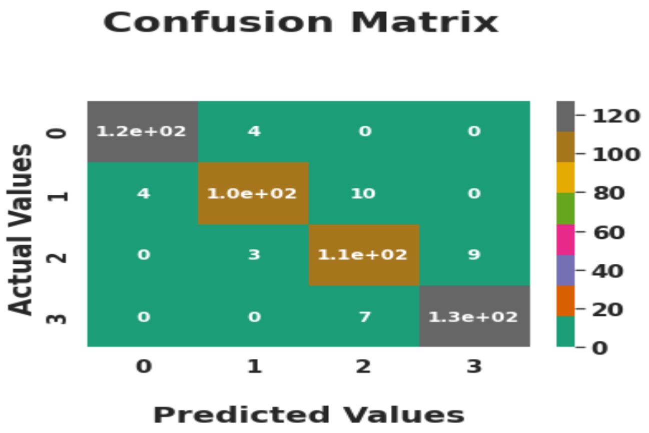
```
precision score:  0.896

recall score:  0.896

              precision    recall  f1-score   support

           0       0.95      0.96      0.96       129
           1       0.88      0.83      0.86       119
           2       0.81      0.85      0.83       118
           3       0.93      0.93      0.93       134

    accuracy                           0.90       500
   macro avg       0.89      0.89      0.89       500
weighted avg       0.90      0.90      0.90       500
```



Hyper Parameter

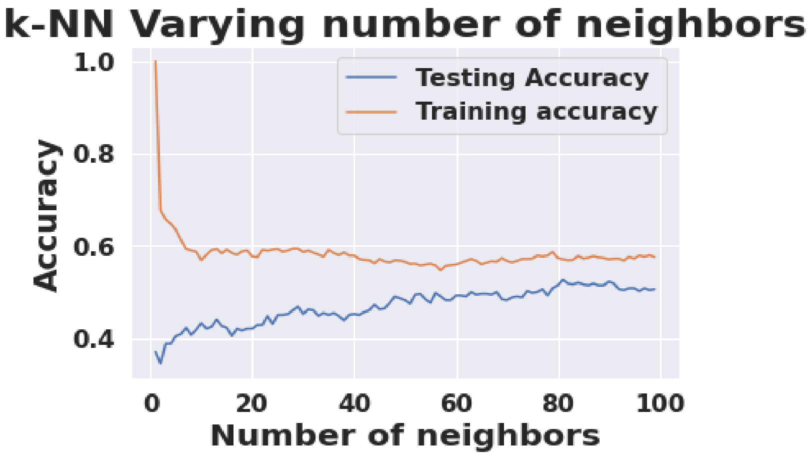


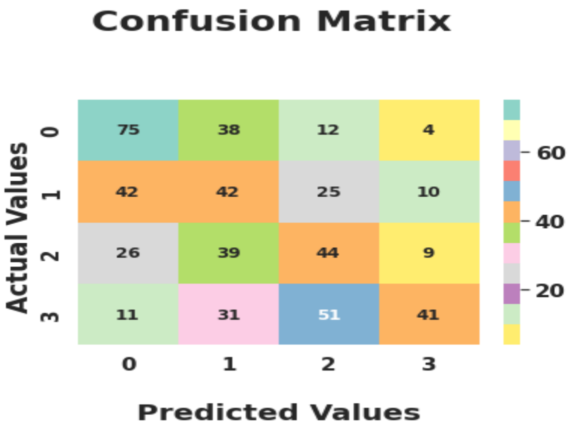
```
precision score:  0.926

recall score:  0.926

              precision    recall  f1-score   support

           0       0.97      0.97      0.97       129
           1       0.94      0.88      0.91       119
           2       0.86      0.90      0.88       118
           3       0.93      0.95      0.94       134

    accuracy                           0.93       500
   macro avg       0.93      0.92      0.92       500
weighted avg       0.93      0.93      0.93       500
```

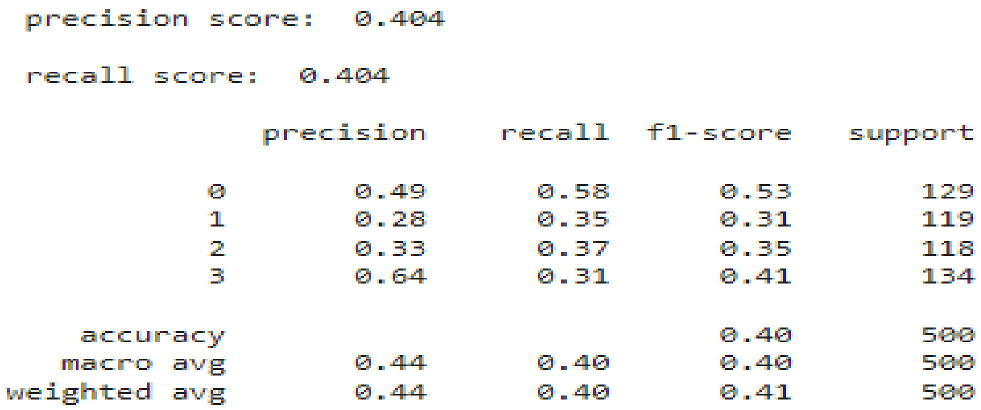


Implementing K_nearest neighbours(knn)

The training score is 63%, and the testing score is 40%.
I got poor grade from this model



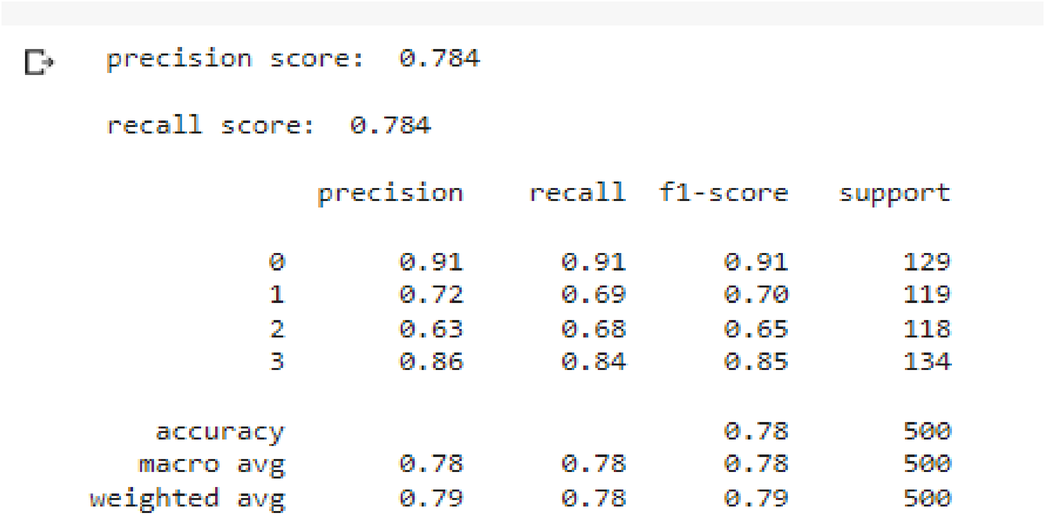




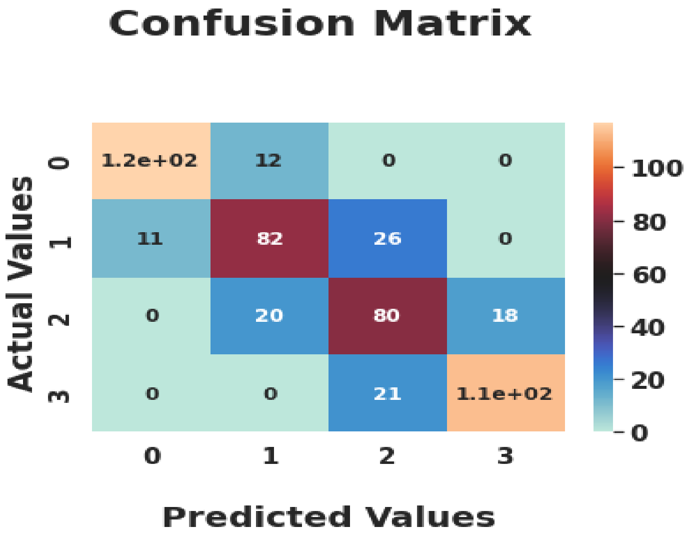
```
precision score:  0.404

recall score:  0.404
              precision    recall  f1-score   support

           0       0.49      0.58      0.53       129
           1       0.28      0.35      0.31       119
           2       0.33      0.37      0.35       118
           3       0.64      0.31      0.41       134

    accuracy                           0.40       500
   macro avg       0.44      0.40      0.40       500
weighted avg       0.44      0.40      0.41       500
```

Implementing Naive Bayes Classifier

For training score is 81% and testing score 78%.
I got poor grade from this model



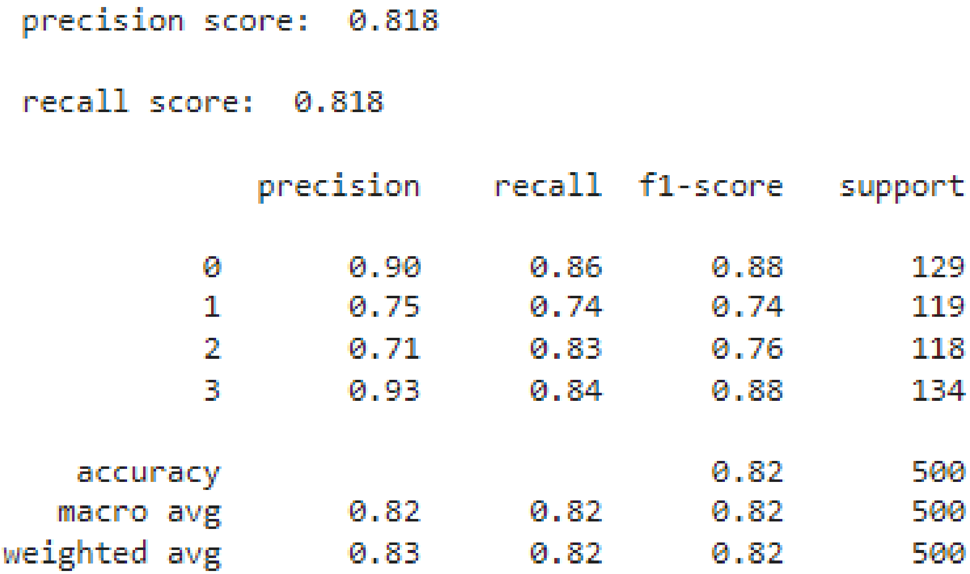
```
precision score:  0.784

recall score:  0.784

              precision    recall  f1-score   support

           0       0.91      0.91      0.91       129
           1       0.72      0.69      0.70       119
           2       0.63      0.68      0.65       118
           3       0.86      0.84      0.85       134

    accuracy                           0.78       500
   macro avg       0.78      0.78      0.78       500
weighted avg       0.79      0.78      0.79       500
```



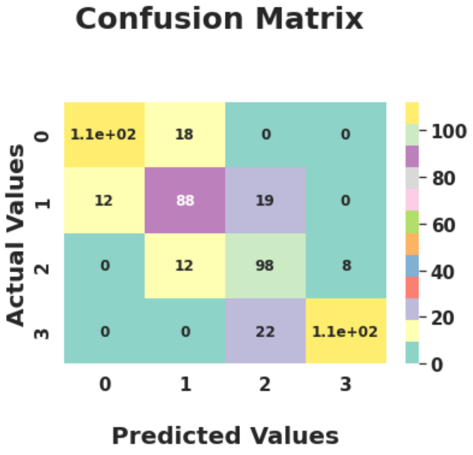
Confusion Matrix

Implementing Support Vector Machine Classifier (SVM)

1.The training score is 95% and the testing score is 81%.
2.I got poor grade from this model



```
precision score:  0.818

recall score:  0.818

              precision    recall  f1-score   support

           0       0.90      0.86      0.88       129
           1       0.75      0.74      0.74       119
           2       0.71      0.83      0.76       118
           3       0.93      0.84      0.88       134

    accuracy                           0.82       500
   macro avg       0.82      0.82      0.82       500
weighted avg       0.83      0.82      0.82       500
```



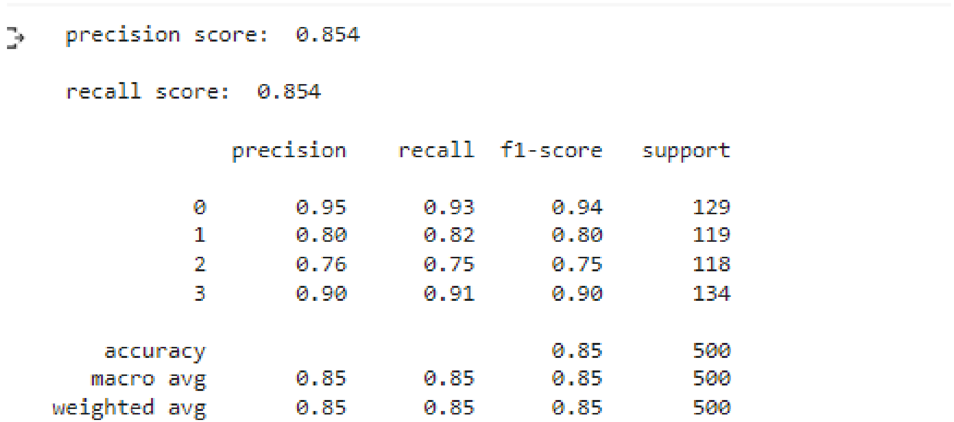
Confusion Matrix

	0	1	2	3
0	1.1e+02	18	0	0
1	12	88	19	0
2	0	12	98	8
3	0	0	22	1.1e+02

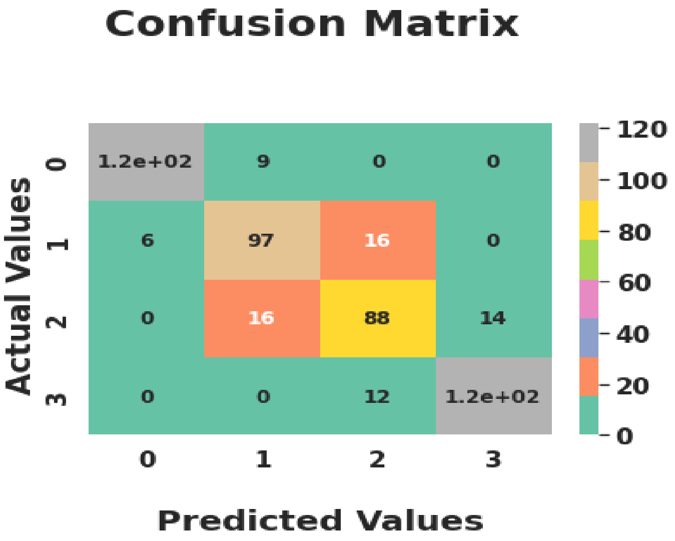
Implementing Decision Tree

1.I did not achieve satisfactory results.

2.The training score is 93%, and the testing score is 85%.

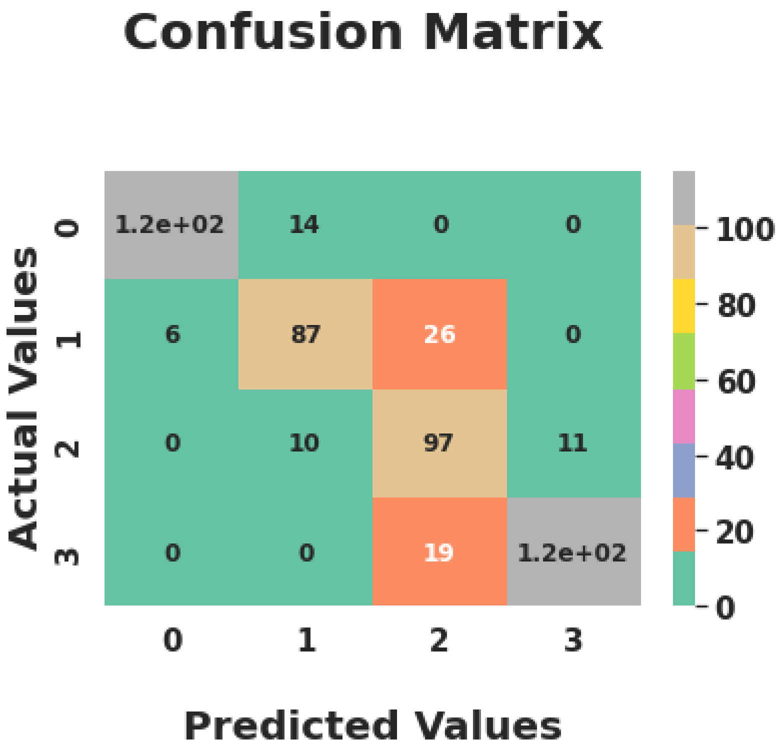


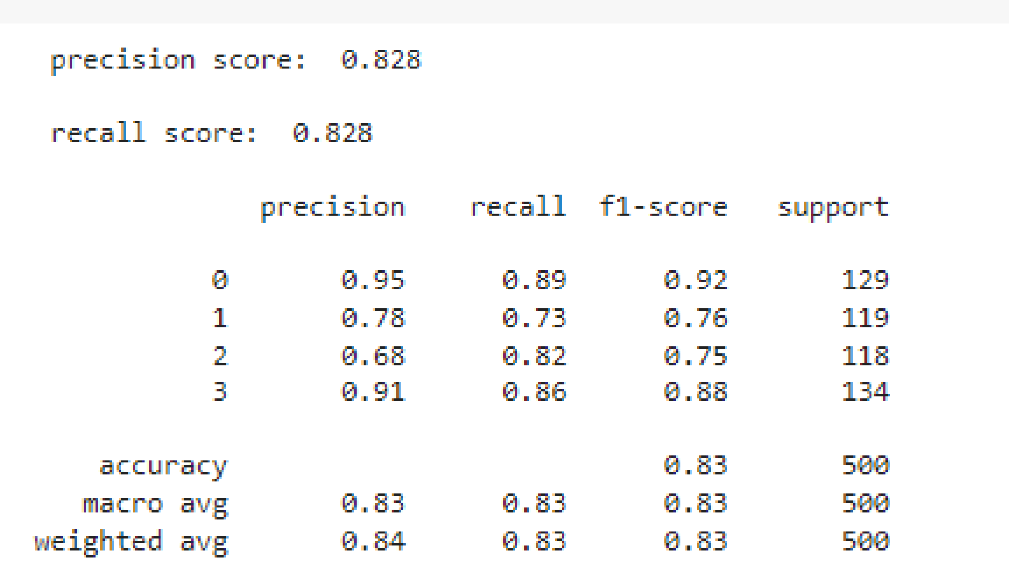
```
precision score:  0.854

recall score:  0.854

              precision    recall  f1-score   support

           0       0.95      0.93      0.94       129
           1       0.80      0.82      0.80       119
           2       0.76      0.75      0.75       118
           3       0.90      0.91      0.90       134

    accuracy                           0.85       500
   macro avg       0.85      0.85      0.85       500
weighted avg       0.85      0.85      0.85       500
```



Cross Validation and Hyper Parameter

The training score is 89%, while the testing score is 82%

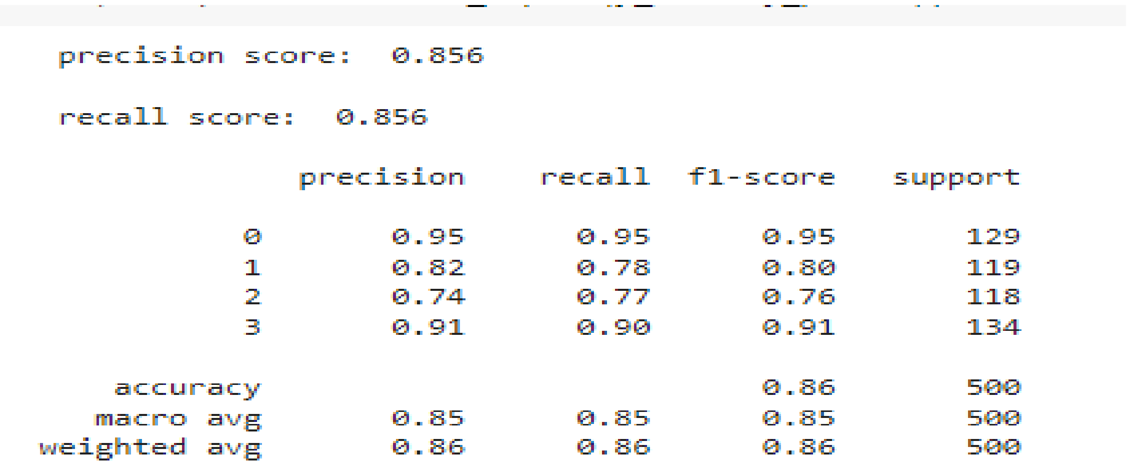




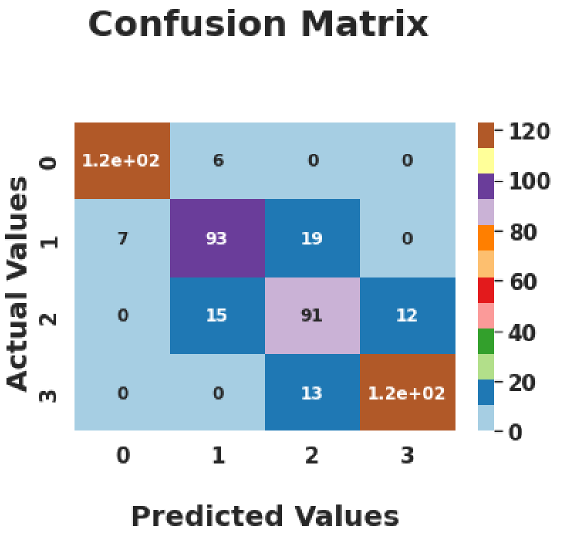
```
precision score:  0.828

recall score:  0.828

              precision    recall  f1-score   support

           0       0.95      0.89      0.92       129
           1       0.78      0.73      0.76       119
           2       0.68      0.82      0.75       118
           3       0.91      0.86      0.88       134

    accuracy                           0.83       500
   macro avg       0.83      0.83      0.83       500
weighted avg       0.84      0.83      0.83       500
```

Implementing Random forest Classifier

1. I did not get Satisfactory result.
2.The training score is 99%, and the testing score is 85%.

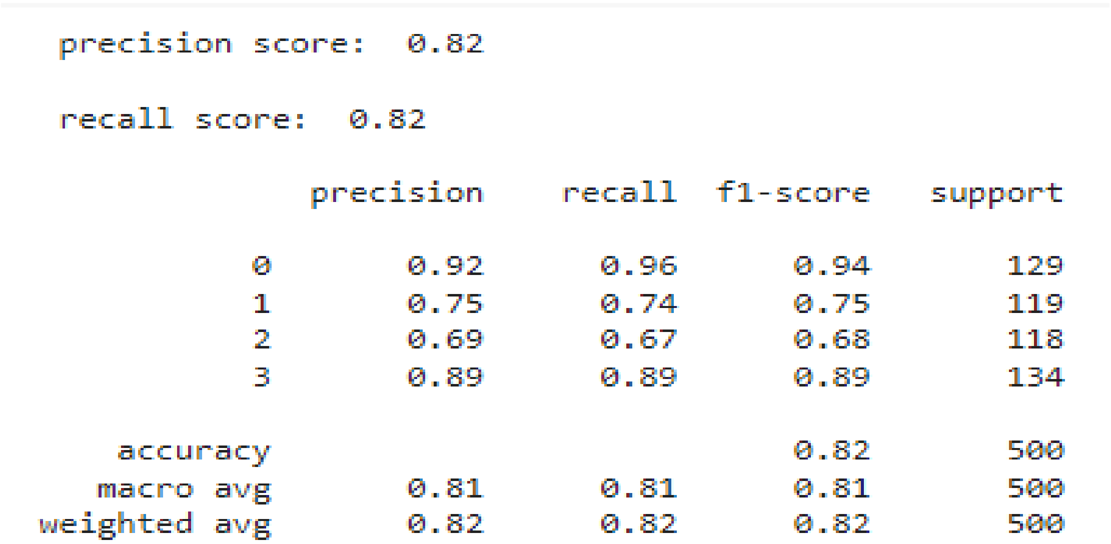


```
    precision score:  0.856

    recall score:  0.856

                    precision    recall  f1-score   support

                0       0.95      0.95      0.95       129
                1       0.82      0.78      0.80       119
                2       0.74      0.77      0.76       118
                3       0.91      0.90      0.91       134

         accuracy                           0.86       500
        macro avg       0.85      0.85      0.85       500
     weighted avg       0.86      0.86      0.86       500
```

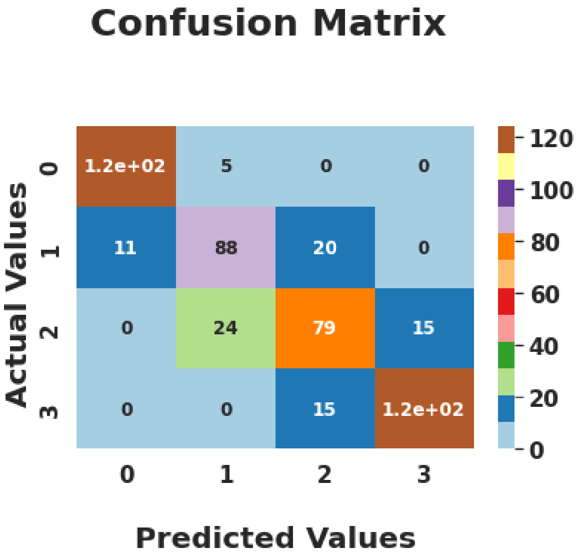


Cross Validation and Hyper Parameter Tuning

- I used the GridsearchCV method for hyperparameter optimization.
- The training score is 94%, while the testing score is 82%.



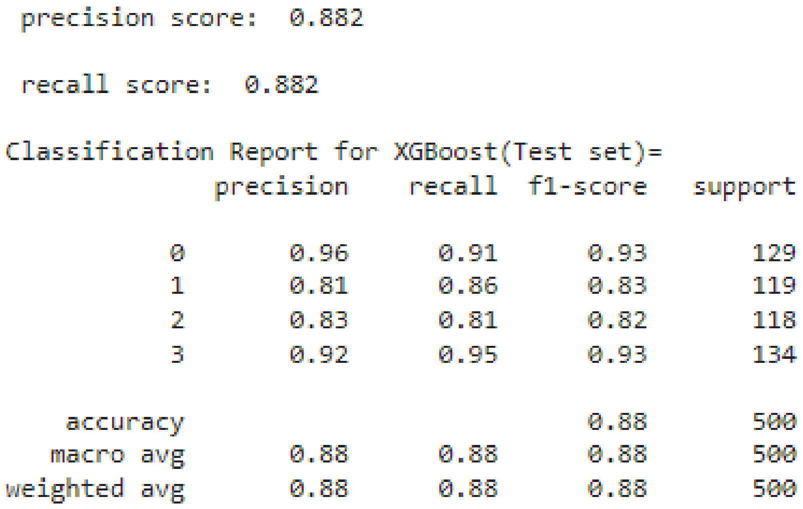
```
  precision score:  0.82

  recall score:  0.82

                precision    recall  f1-score   support

             0       0.92      0.96      0.94       129
             1       0.75      0.74      0.75       119
             2       0.69      0.67      0.68       118
             3       0.89      0.89      0.89       134

      accuracy                           0.82       500
     macro avg       0.81      0.81      0.81       500
  weighted avg       0.82      0.82      0.82       500
```



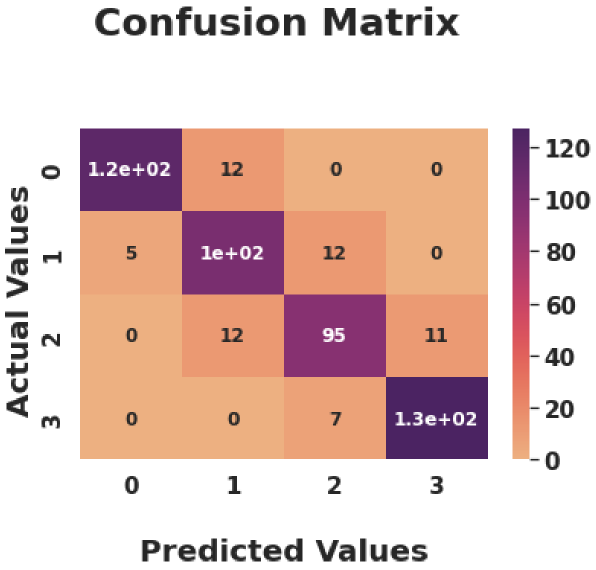
Confusion Matrix

Implementing XGB Classifier

- ✓ I received pretty good results.
- ✓ The testing score is 88% and the training score is 100%.



```
 precision score:  0.882

 recall score:  0.882

Classification Report for XGBoost(Test set)=
              precision    recall  f1-score   support

           0       0.96      0.91      0.93       129
           1       0.81      0.86      0.83       119
           2       0.83      0.81      0.82       118
           3       0.92      0.95      0.93       134

    accuracy                           0.88       500
   macro avg       0.88      0.88      0.88       500
weighted avg       0.88      0.88      0.88       500
```

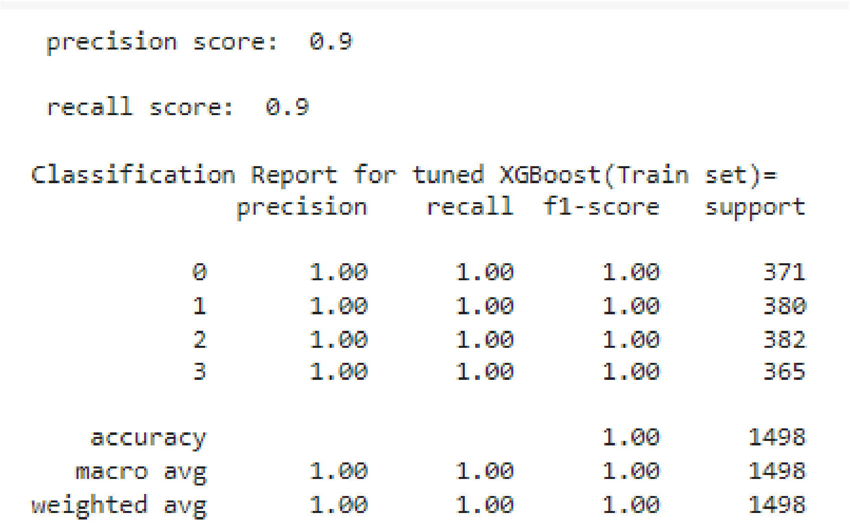


Confusion Matrix

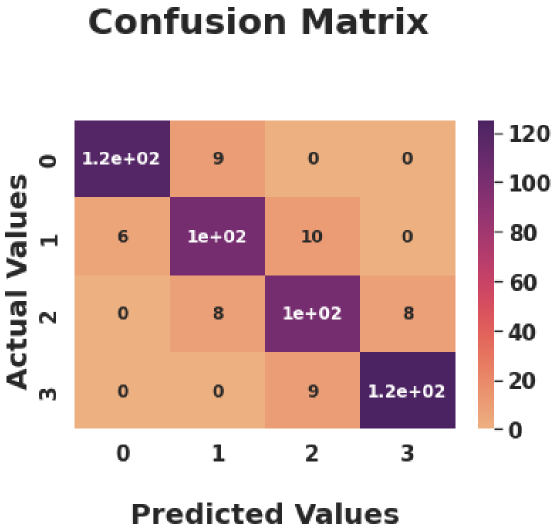
	0	1	2	3
0	1.2e+02	12	0	0
1	5	1e+02	12	0
2	0	12	95	11
3	0	0	7	1.3e+02

Cross Validation and Hyper Parameter Tuning

- ✓ The cross validation of the XGB classifier produced positive results!
- ✓ Test scores is 90% and training is 98% respectively.



```
  precision score:  0.9

  recall score:  0.9

 Classification Report for tuned XGBoost(Train set)=
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       371
           1       1.00      1.00      1.00       380
           2       1.00      1.00      1.00       382
           3       1.00      1.00      1.00       365

    accuracy                           1.00      1498
   macro avg       1.00      1.00      1.00      1498
weighted avg       1.00      1.00      1.00      1498
```



Conclusion

- ✓ From EDA, we can observe that there are four different price categories for mobile phones. Almost the same number of components are present.
- ✓ Half of the devices have Bluetooth, whereas the other half do not.
- ✓ When the price range widens, the battery gradually gets more powerful.
- ✓ The price range for Ram increases steadily as it moves from low to extremely high prices.
- ✓ Expensive phones are smaller.
- ✓ The most important factors in determining the price range of a mobile phone were RAM, battery life, and pixels.
- ✓ Based on the results of the aforementioned trials, we can say that the XGboost classifier and Logistic regression with the use of hyperparameters produced the best outcomes.

THANK YOU