Project 11: Mobile Price Prediction

Name:	**Sugandh Mittal, Sumouli Chakraborty**
Registration No./Roll No.:	20278, 20279
Institute/University Name:	IISER Bhopal
Program/Stream:	EECS
Problem Release date:	February 02, 2023
Date of Submission:	March 01, 2023

1 Data Description

The data was provided by our professor.It contains 2000 data points, 20 features ranging from blue(bluetooth),fc(front camera),pc(primary camera)etc. and 4 classes - 0(cheap),1(moderate),2 (economical) and 3(expensive).

2 Introduction

Aim: Predicting the price range of mobile phones based on their specifications.
Task: Classify the price range into four categories. We will discuss detailed data engineering,model selection,and performance evaluation.

2.1 Descriptive Data Analysis

2.1.1 data.head

We get to know about the features present in the training data.

	battery_power	blue	clock_speed	dual_sim	fc	four_g	int_memory	m_dep	mobile_wt	n_cores	...	px_height
0	842	0	2.2	0	1	0	7	0.6	188	2	...	20
1	1021	1	0.5	1	0	1	53	0.7	136	3	...	905
2	563	1	0.5	1	2	1	41	0.9	145	5	...	1263
3	615	1	2.5	0	0	0	10	0.8	131	6	...	1216
4	1821	1	1.2	0	13	1	44	0.6	141	2	...	1208

Figure 1: Name of Features

2.1.2 data.info

Get information about number of columns,column labels,column data types,memory usage,range index,and the number of cells in each column (non-null values).Our data is numerical with no missing values.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 21 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   battery_power  2000 non-null   int64
 1   blue           2000 non-null   int64
 2   clock_speed    2000 non-null   float64
 3   dual_sim       2000 non-null   int64
 4   fc             2000 non-null   int64
 5   four_g         2000 non-null   int64
 6   int_memory     2000 non-null   int64
 7   m_dep          2000 non-null   float64
 8   mobile_wt      2000 non-null   int64
 9   n_cores        2000 non-null   int64
 10  pc             2000 non-null   int64
 11  px_height      2000 non-null   int64
 12  px_width       2000 non-null   int64
 13  ram            2000 non-null   int64
 14  sc_h           2000 non-null   int64
 15  sc_w           2000 non-null   int64
 16  talk_time      2000 non-null   int64
 17  three_g        2000 non-null   int64
 18  touch_screen   2000 non-null   int64
 19  wifi           2000 non-null   int64
 20  price_range    2000 non-null   int64
dtypes: float64(2), int64(19)
memory usage: 328.2 KB
```

Figure 2: Information about Data

2.1.3 data.describe

Describe about mean,median,standard deviation,maximum value,25,50 and 75 percentile for each column present in training data.Few features are shown as examples.

	battery_power	blue	clock_speed	dual_sim	fc	four_g
count	2000.000000	2000.0000	2000.000000	2000.000000	2000.000000	2000.000000
mean	1238.518500	0.4950	1.522250	0.509500	4.309500	0.521500
std	439.418206	0.5001	0.816004	0.500035	4.341444	0.499662
min	501.000000	0.0000	0.500000	0.000000	0.000000	0.000000
25%	851.750000	0.0000	0.700000	0.000000	1.000000	0.000000
50%	1226.000000	0.0000	1.500000	1.000000	3.000000	1.000000
75%	1615.250000	1.0000	2.200000	1.000000	7.000000	1.000000
max	1998.000000	1.0000	3.000000	1.000000	19.000000	1.000000

Figure 3: Description of Data

2.1.4 data.shape

Gives us information about the no. of rows vs no. of columns.We get 2000X21 i.e 20 features and 1 column of target variables.

2.2 Exploratory Data Analysis

2.2.1 Bluetooth vs Count

This plot shows half the devices have bluetooth,and half don't.

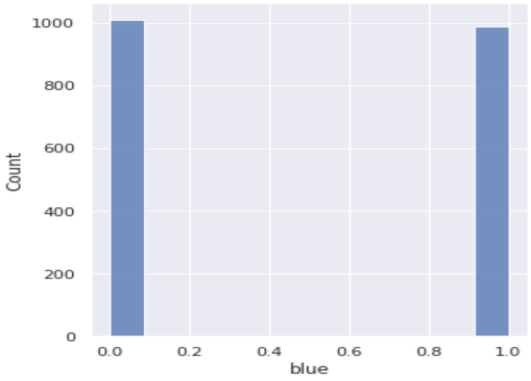

Figure 4: Bluetooth Count Plot

2.2.2 Percent of phones supporting 4G

This plot shows 52.1 percent mobiles support 4G.

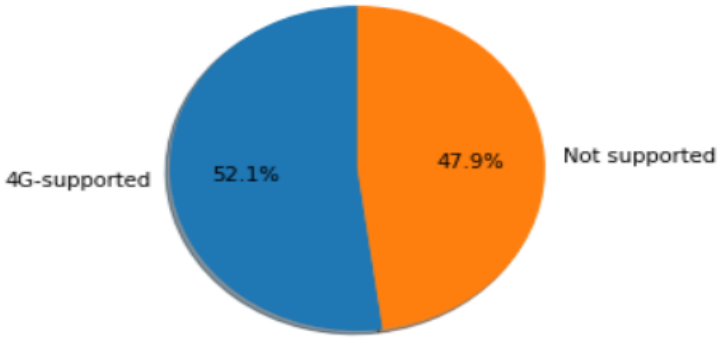

Figure 5: Phones supporting 4G

2.2.3 Price Range vs Talk-Time

It shows talk time in different price range.

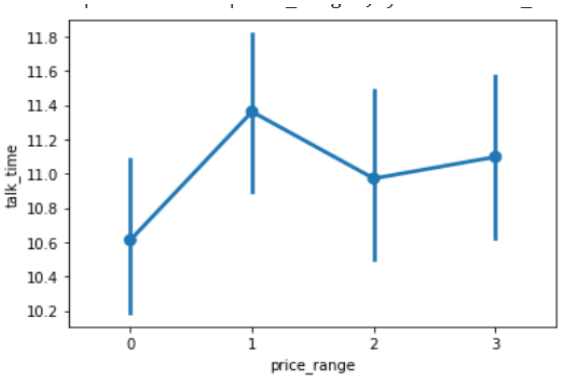

Figure 6: Price Range vs Talk-Time

2.2.4 RAM vs Price Range

This plot gives density distribution for ram vs price range.

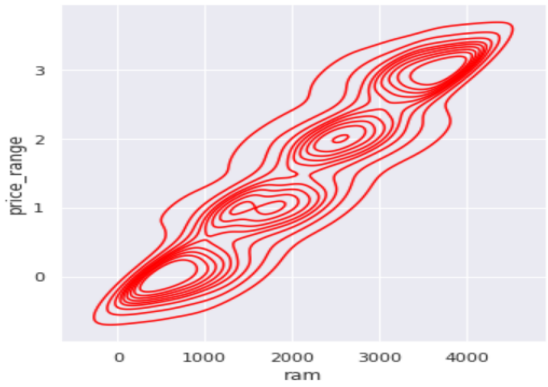

Figure 7: RAM vs Price Range

We also found out correlation matrix.Rest of the plots are in code file.

3 Methods

For Phase-I of the project we explored two classification models: Random Forest Classification model, Gaussian Naive Bayes model along with hyperparameter tuning,tried to compare the outputs to optimize performances of the models to see which one is performing best.
For hyperparameter tuning we can use Grid search,Random Search,Bayesian optimisation etc.

4 Evaluation Criteria:

We compared the performance of the above two models on certain criterias like- Accuracy,Precision,Recall,F-1 Score,Macro and Micro averaging techniques.

5 Analysis of Results:

In this section we compare accuracy, recall,precision,f-1 score of different classifier models prepared. The following figures9 and 11 shows the output and confusion matrix of Random forest classifier and Gaussian Naive Bayes classifier before and after parameter tuning which generally gives increased accuracy scores. So from this we see that using different values for the parameters than the default one,gives better values for for evaluation criteria used.

Random Forest Classifier Accuracy Score: 0.88

	precision	recall	f1-score	support
0	1.00	0.92	0.96	109
1	0.77	0.93	0.84	89
2	0.85	0.77	0.81	106
3	0.92	0.91	0.91	96
accuracy			0.88	400
macro avg	0.88	0.88	0.88	400
weighted avg	0.89	0.88	0.88	400

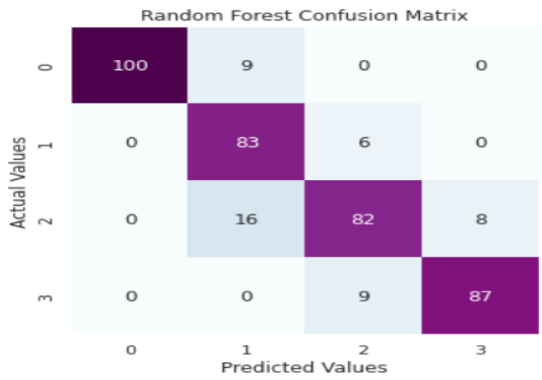

Figure 8: Output of Random Forest Classifier before parameter tuning

Random Forest Classifier Accuracy Score: 0.91

	precision	recall	f1-score	support
0	0.99	0.95	0.97	109
1	0.87	0.96	0.91	89
2	0.87	0.84	0.86	106
3	0.91	0.90	0.90	96
accuracy			0.91	400
macro avg	0.91	0.91	0.91	400
weighted avg	0.91	0.91	0.91	400

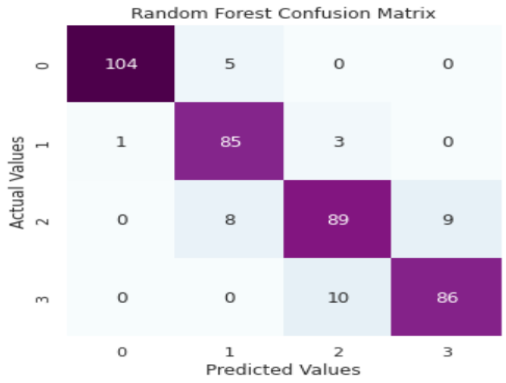

Figure 9: Output of Random Forest Classifier

Accuracy Score: 0.8

	precision	recall	f1-score	support
0	0.96	0.89	0.92	109
1	0.65	0.76	0.70	89
2	0.70	0.69	0.70	106
3	0.91	0.85	0.88	96
accuracy			0.80	400
macro avg	0.81	0.80	0.80	400
weighted avg	0.81	0.80	0.80	400

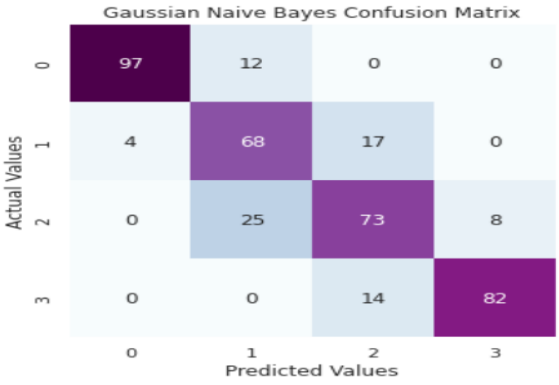

Figure 10: Output of Gaussian Naive Bayes Classifier before Parameter Tuning

Accuracy Score: 0.7975

	precision	recall	f1-score	support
0	0.96	0.88	0.92	109
1	0.64	0.75	0.69	89
2	0.70	0.69	0.70	106
3	0.91	0.86	0.89	96
accuracy			0.80	400
macro avg	0.80	0.80	0.80	400
weighted avg	0.81	0.80	0.80	400

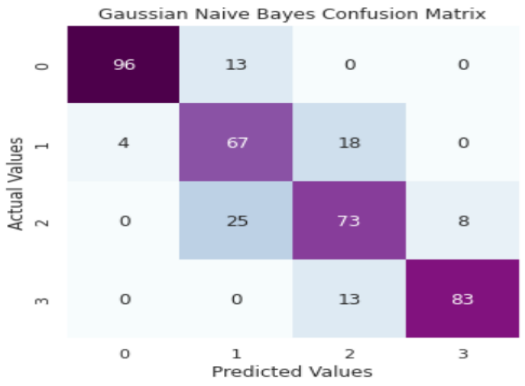

Figure 11: Output of Gaussian Naive Bayes Classifier

6 Discussions and Conclusion

We started with data analysis,splitting it,followed by pre-processing.Made Random Forest and Gaussian Naive-Bayes models, discussed the effect of hyper-parameter tuning on their accuracy scores. In phase-2 of the project,we'll explore further models to find best suitable classifier for this problem.

7 Roles Played

1.Sugandh Mittal- Code-Half the data analysis and one model, Report- section 1,2,4
2.Sumouli Chakraborty - Code-Rest half of the data analysis,one model, Report- section 3,5,6.