Mobile Price Prediction

Prince Jain, Vikas Shrivas,
Rishabh Patidar
Data science trainees,
Almabetter, Bangalore

Abstract:

To predict "If the mobile with given features will be **Economical or Expensive**" is the main motive of this research work. Different feature selection algorithms are used to identify and remove less important and redundant features and have minimum computational complexity. Different classifiers are used to achieve as higher accuracy as possible. Results are compared in terms of highest accuracy achieved and minimum features selected. Conclusion is made on the base of best feature selection algorithm and best classifier for the given dataset. This work can be used in any type of marketing and business to find optimal product (with minimum cost and maximum features). Future work is suggested to extend this research and find more sophisticated solution to the given problem and more accurate tool for price estimation.

1. Problem Statement:

Mobile phones come in all sorts of prices, features, specifications and all. Price estimation and prediction is an important part of consumer strategy. Deciding on the correct price of a product is very important for the market success of a product. A new product that has to be launched, must have the correct price so that consumers find it appropriate to buy the product.

The data contains information regarding mobile phone features, specifications etc. and their price range. The various features and information can be used to predict the price range of a mobile phone.

2. Analysis of Mobile Price Prediction:

The dataset was derived from market research. And this data consists 21 columns and 2000 rows which have contains various specification about mobile phone and the following data attributes are given which are: -

- "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
- 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)

We remove columns names Sc_h and Sc_w. And we have to predict the price of mobile phone using various parameters.

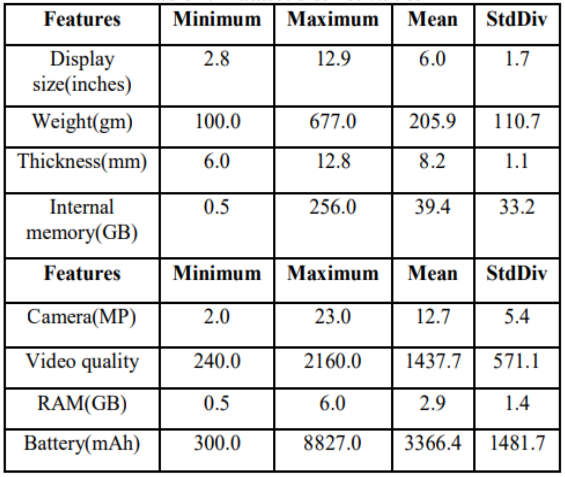
3. Research Methods:

In this paper, we propose a classification framework based on ensemble learning to classify and predict price for the mobile phone. The framework involved four steps, including data collection and understanding, data preprocessing, data cleaning and Exploratory Data Analysis (EDA), Machine learning models.**3.1. Data collection and understanding:** The data primarily contained the following attributes of information: Battery power, Blue, Clock speed, Dual_sim, 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, 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).

3.2. Data Preprocessing: In the dataset, we go through the 20 columns and perform various operations and obtains the valuable outcomes and insights. Training on these attributes would not only increase the required training time but also render the training results unreasonable or impractical; therefore, data preprocessing operations are essential. At this stage, the mobile price prediction dataset was processed through data cleaning, feature engineering, and data normalization.

3.3. Data cleaning: Data cleaning aims to reduce the dimensions of the GTD dataset by detecting and deleting irrelevant or redundant attributes and case records.

Table 1. Dataset distinct values



Features	Minimum	Maximum	Mean	StdDiv
Display size(inches)	2.8	12.9	6.0	1.7
Weight(gm)	100.0	677.0	205.9	110.7
Thickness(mm)	6.0	12.8	8.2	1.1
Internal memory(GB)	0.5	256.0	39.4	33.2
Features	**Minimum**	**Maximum**	**Mean**	**StdDiv**
Camera(MP)	2.0	23.0	12.7	5.4
Video quality	240.0	2160.0	1437.7	571.1
RAM(GB)	0.5	6.0	2.9	1.4
Battery(mAh)	300.0	8827.0	3366.4	1481.7

3.4. Exploratory Data Analysis (EDA): Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods.

3.5 Machine learning model: Various classification models are follows:

- K-Nearest Neighbors.
- Support Vector Machines.
- Decision Tree Classifiers/Random Forests.
- Naive Bayes.
- Logistic Regression.

4. Datasets:

Mobile phones come in all sorts of prices, features, specifications and all. Price estimation and prediction is an important part of consumer strategy. Deciding on the correct price of a product is very important for the market success of a product. A new product that has to be launched, must have the correct price so that consumers find it appropriate to buy the product.

The data contains information regarding mobile phone features, specifications etc. and their price range. The various features and information can be used to predict the price range of a mobile phone.

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 (e.g.: - RAM, Internal Memory, etc.) and its selling price. In this problem, we do not have to predict the actual price but a price range indicating how high the price.

5. Analysis:

This section consists of details regarding the visual results:

5.1 Different types of parameters:

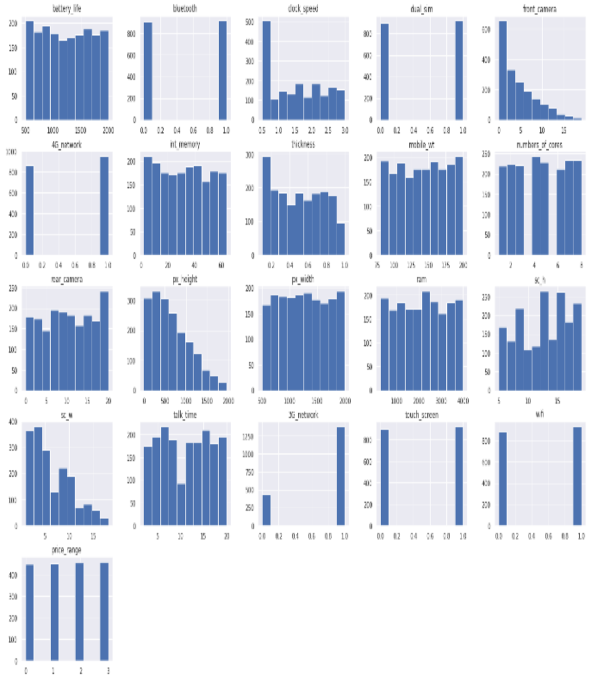
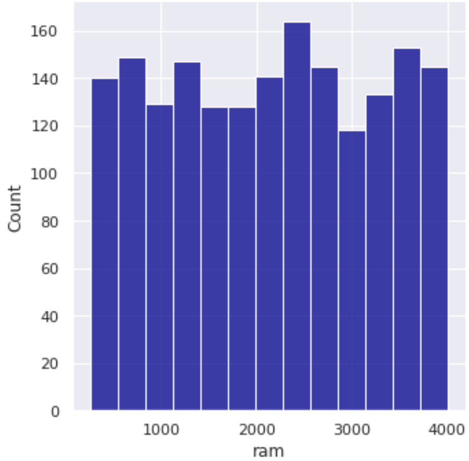


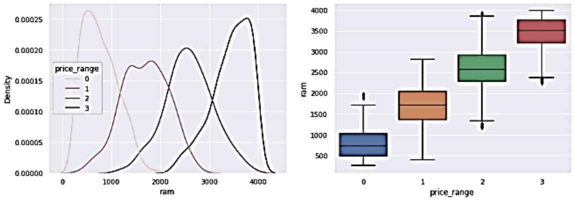
Figure 1: All the parameters bar charts are shown

In this fig. we see the various parameter which are useful in determining the dependent variable (price). There are 20 variables are present.

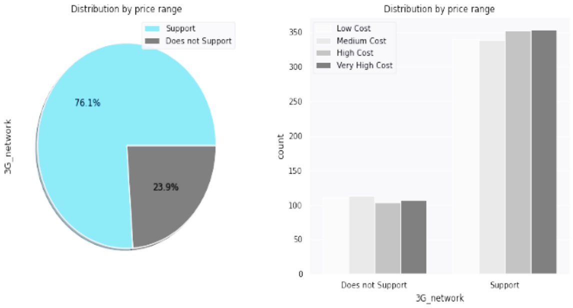
5.1.1 RAM:

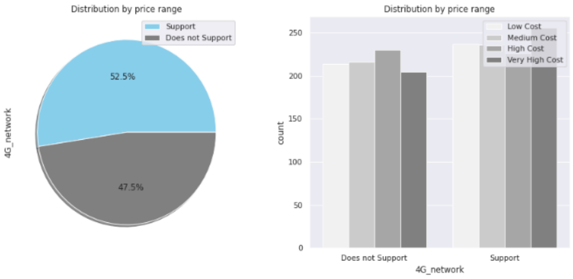


RAM are the important parameter for the prediction in the price and the chart are shows that the variation in the ram is 256mb to 4gb and as the ram increases the price is also increases.

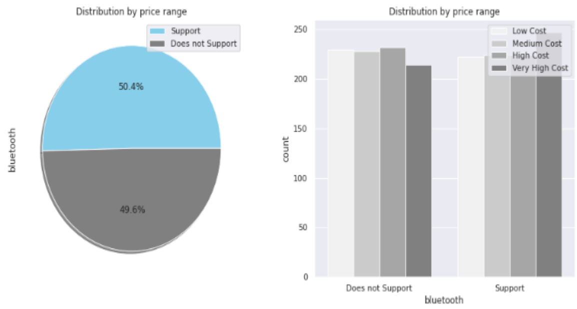


5.1.2 Connectivity:





A study by **Mobile Price Range Prediction** has shown that 76.1% of mobile phone users have 3G connectivity, while 23.9% do not support 3G. There are 52.5% of mobile phones with 4G connectivity and 43.5% do not support 4G. According to TRAI data, 50.4% of the users have Bluetooth connectivity and 49.6% do not have it. All these mobile phones are expensive, though.



5.3 Camera in mega pixels

Rear camera

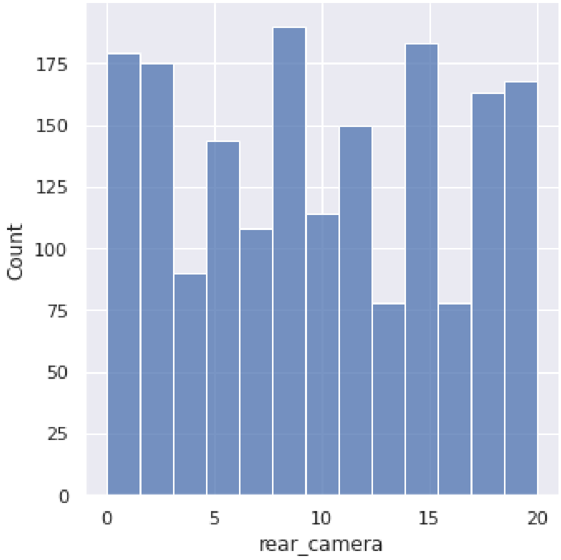
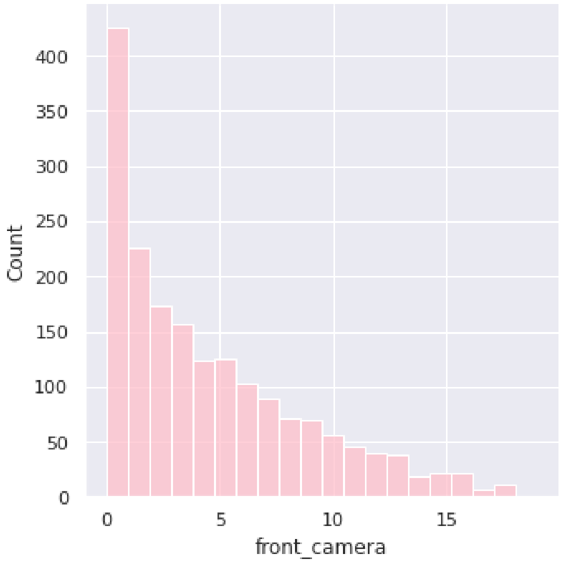


Figure4: rear camera

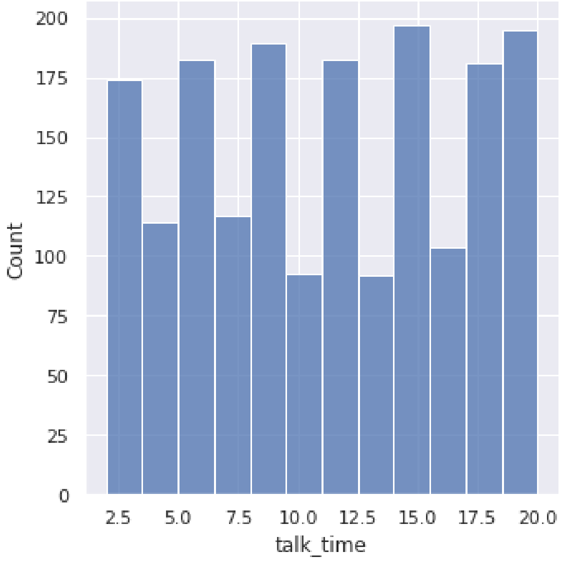
In this bar chart, we found that some mobile cameras do not contain a camera, and show zero. We also saw that the maximum mobile camera is 8MP and after it is 13MP.

Front camera:



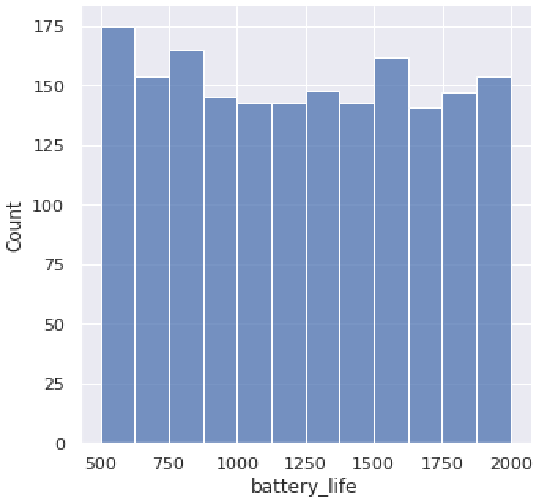
The majority of phones do not contain a front-facing camera, and the maximum number of phones currently on sale contains a 2mp camera.

5.4Talktime:



The bar chart above shows the range of talk time among the phones. The lowest range is 2.5 hours, and the longest is 20 hours.

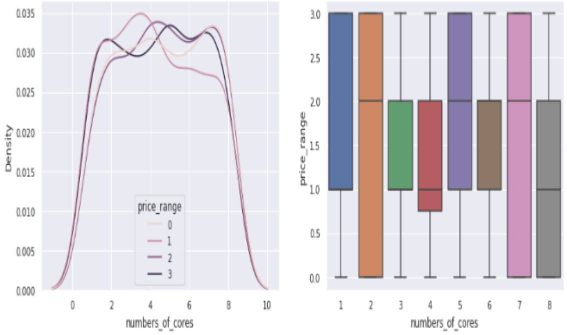
5.5 Battery in MAH:



The battery life of a mobile phone typically ranges from 500 to 2000 mAh. The most frequently purchased mobile phone with a 500 mAh battery is the most popular.

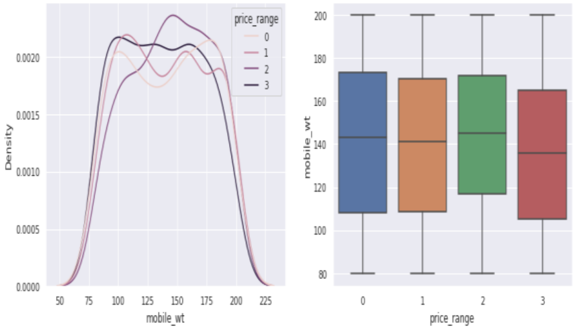
5.6 Discuss various parameters and their relationship to price

5.6.1Numbers of core vs price



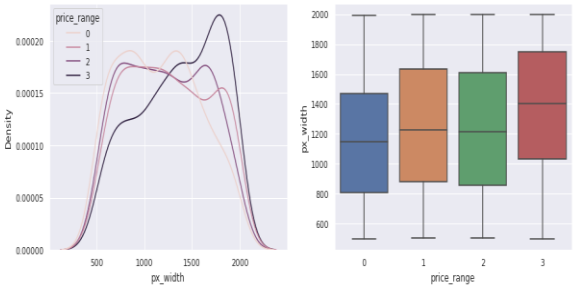
The above chart shows the number of core 2 and 7 available in the price range 0 to 3, whereas we saw that the number of course 8 is not available at a high price, and core number 1 is not available at a low price.

5.6.2 Weight vs price



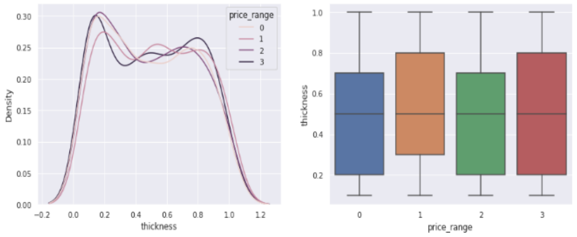
It is observed that the cost of a phone is directly proportional to its weight. The price of a cell phone rises as the weight of the phone decreases.

5.6.3 Pixel Resolution Width



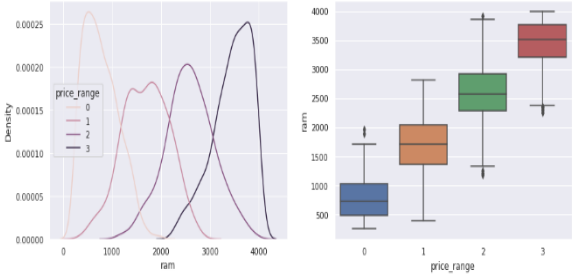
As we move from Low cost to Very high cost, the pixel widths of mobiles do not increase in absolute terms. However, mobile with 'Medium cost' and 'High cost' has almost equal pixel widths so we can say that it would be a driving factor in deciding price range.

5.6.6 Thickness vs price



Thick phones are available at the lowest price. The thickness of mobiles ranges from 0.2 cm to almost one centimetre thick. The cost of thick phones is low or may be high.

5.6.7 Ram vs price



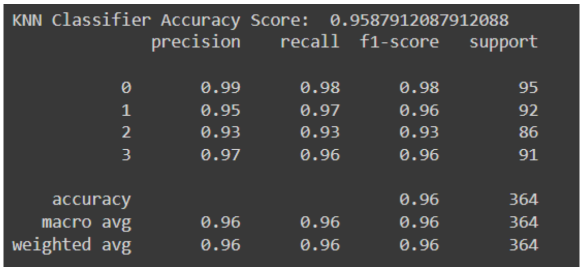
The higher the RAM, the more expensive a smartphone is likely to be.

RAM ∝ MOBILE PRICE

6. Machine learning models:

6.1 K-Nearest Neighbors

The KNN method saves all available data and classifies a new data point based on its similarity to the existing data. This implies that fresh data may be quickly sorted into a welldefined category using the KNN method **Accuracy_ Score: 95.87%**

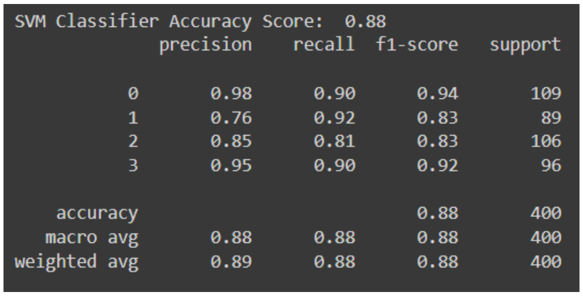


```
KNN Classifier Accuracy Score:  0.9587912087912088
              precision    recall  f1-score   support

           0       0.99      0.98      0.98        95
           1       0.95      0.97      0.96        92
           2       0.93      0.93      0.93        86
           3       0.97      0.96      0.96        91

    accuracy                           0.96       364
   macro avg       0.96      0.96      0.96       364
weighted avg       0.96      0.96      0.96       364
```

6.2 Support Vector Machines: The linear SVM model had a **classification accuracy of 88%** with those transcript variables, four fewer variables than logistic regression

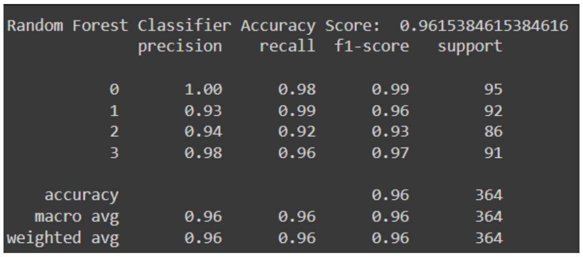


```
SVM Classifier Accuracy Score:  0.88
              precision    recall  f1-score   support

           0       0.98      0.90      0.94       109
           1       0.76      0.92      0.83        89
           2       0.85      0.81      0.83       106
           3       0.95      0.90      0.92        96

    accuracy                           0.88       400
   macro avg       0.88      0.88      0.88       400
weighted avg       0.89      0.88      0.88       400
```

6.3 Random Forests: A random forest is built on a variety of decision trees. Every decision tree is made up of nodes that represent decisions, leaf nodes and a root node. The leaf nodes of each tree represent the decisions in the decision tree, and the root node represents the final result of that decision tree. The final product can be determined using a majority-voting procedure. Let us now implement our random forest algorithm. **Accuracy_ Score: 96.15%**

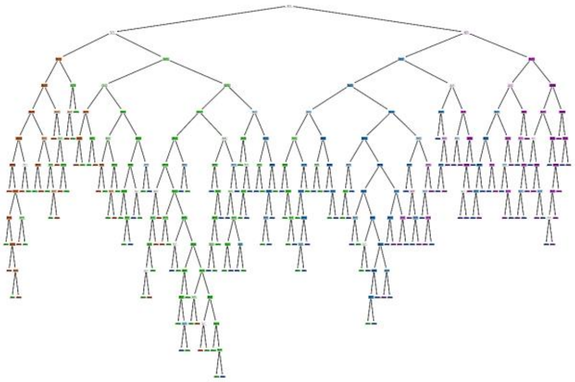


```
Random Forest Classifier Accuracy Score:  0.9615384615384616
              precision    recall  f1-score   support

           0       1.00      0.98      0.99        95
           1       0.93      0.99      0.96        92
           2       0.94      0.92      0.93        86
           3       0.98      0.96      0.97        91

    accuracy                           0.96       364
   macro avg       0.96      0.96      0.96       364
weighted avg       0.96      0.96      0.96       364
```

6.5 Decision Tree Classifiers:

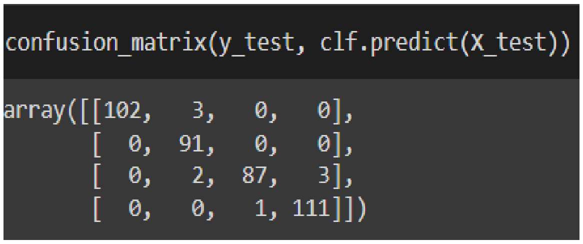


The model overfit because the training dataset accuracy was 100% for all 5 different folds of the cross validation, **while the average testing dataset accuracy was 83.14%.** An 83.14% accuracy for the testing set is pretty good, but I believe this model could do better. I will try to tune this model by adding more layers with smaller number of neurons per layer.

6.6 Naive Bayes: By comparing the actual and predicted values of the **Naïve Bayes method, an accuracy of 84.61% was achieved**.

6.7 Logistic Regression: Overall, some good baseline statistics. I will use KFold Cross Validation to ensure that the model is not overfitting and get a more realistic accuracy for the training and testing datasets. For future scenarios, I will use 5 folds for KFold Cross Validation. Note that Stratified Fold Cross Validation isn't needed as the price range data is spread out (equal number of 0, 1, 2, and 3)

Train_accuracy: 97.71%
Test_accuracy: 96.25%



```
confusion_matrix(y_test, clf.predict(X_test))
```

```
array([[102,   3,   0,   0],
       [  0,  91,   0,   0],
       [  0,   2,  87,   3],
       [  0,   0,   1, 111]])
```

7. Technologies used:

Python: Python is a high-level interpreted language that supports different platforms like Windows, Linux, Mac, Raspberry Pi, etc. Python can be used for creating web applications, database systems, handle big data, perform complex mathematical calculations. Python can be treated in an object-oriented, functional or procedural way.

Google Colab: Collaboratory, or "Colab" for short, is a product from Google Research. Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education. More technically, Colab is a hosted Jupyter notebook service that requires no setup to use, while providing access free of charge to computing resources including GPUs.

Python packages: Following are some of the python packages used in this project.

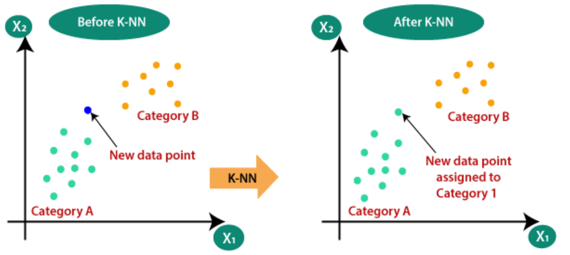
Mathplotlib: Matplotlib is a visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

Pandas: Pandas is an open-source library that is made mainly for working with relational or labeled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series. This library is built on top of the NumPy library. Pandas is fast and it has high performance & productivity for users.

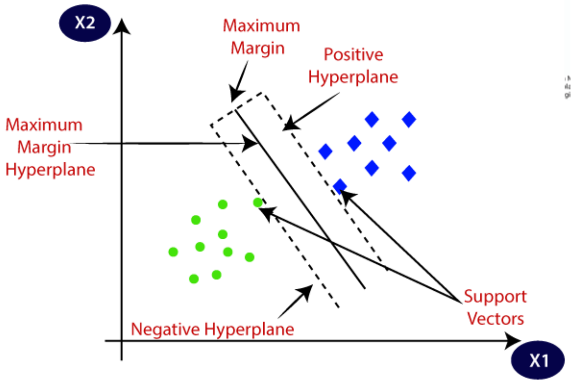
NumPy: It provides structures for multiple dimensional array objects and tools for related operations. NumPy is usually used for high performance scientific computational tasks.

Seaborn: Seaborn is an visualization library for statistical graphics plotting in Python. It provides default styles and color palettes to make statistical plots more attractive. It is built on the top of matplotlib library and also closely integrated to the data structures from pandas. Seaborn aims to make visualization the central part of exploring and understanding data. It provides dataset-oriented APIs, so that we can switch between different visual representations for same variables for better understanding of dataset**.**

k-nearest neighbors': The k-nearest neighbors' algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.



Support Vector Machine (SVM): It is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well its best suited for classification. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number of features exceeds three.



Decision tree regression: Observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.

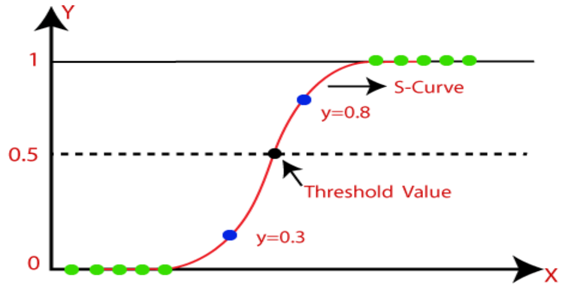
Random Forest Regression: It is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

Naïve Bayes: Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

The Naive Bayes classifier works **on the principle of conditional probability, as given by the Bayes theorem**. While calculating the math on probability, we usually denote probability as P. Some of the probabilities in this event would be as follows: The probability of getting two heads = 1/4.

Logistic regression: Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.



8. Conclusion:

- Classifiers are a set of mathematical algorithms that organize data into groups. They are used for problem-solving, decision-making, and marketing activities. Classifier designers must consider many factors when creating an algorithm, including the nature of the data being analyzed, as well as the goals of the classifier's user.
- The following chart shows a comparison of mobile phones in four price ranges. The number of elements is almost similar. Half the devices have Bluetooth and half don't. There is a gradual increase in battery capacity as the price range increases. Ram

has continuous increases with price ranges while moving from Low cost to Very high cost. Costly phones are lighter. RAM, battery power, and pixels played more significant role in deciding the price range of mobile phones. From all the above experiments, we can conclude that **Logistic regression, k-nearest neighbors, Support vector machines, Decision tree classifiers/random forests and Naive bayes** gave us the best results.

References:

1. Google.com
2. GeeksforGeeks.
3. ResearchGate.
4. https://www.javatpoint.com/machine-learning-naive-bayes-classifier.
5. https://colab.research.google.com/drive/1VJslHcwQi-tHayq3IVxi3iHr0m86p0QF#scrollTo=NvXGvkp8TIKG