**Capstone Project**

**Mobile Price Range Prediction**

**Suraj Kad**

**(**[**suraj.kad.90@gmail.com**](mailto:suraj.kad.90@gmail.com)**)**

**Shubham Sawant (**[**shubhamsawant248@gmail.com**](mailto:shubhamsawant248@gmail.com)**)**

**Tanvir Patel**

**(**[**tanviripatel1998@gmail.com**](mailto:tanviripatel1998@gmail.com)

**DATA SCIENCE TRAINEES,**

**TEAM ALMABETTER**

****

## **Introduction:**

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.

**Problem Statement:**

# 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. In this problem, we do not have to predict the actual price but a price range indicating how high the price is.

**The data features are as follows:**

* Battery Power in mAh
* Has BlueTooth or not
* Microprocessor clock speed
* The phone has dual sim support or not
* Front Camera Megapixels
* Has 4G support or not
* Internal Memory in GigaBytes
* Mobile Depth in Cm
* Weight of Mobile Phone
* Number of cores in the processor
* Primary Camera Megapixels
* Pixel Resolution height
* Pixel resolution width
* RAM in MB
* Mobile screen height in cm
* Mobile screen width in cm
* Longest time after a single charge
* 3g or not
* Has touch screen or not
* Has wifi or not

## **Methodology:**

We will proceed with reading the data, and then perform data analysis. The practice of examining data using analytical or statistical methods in order to identify meaningful information is known as data analysis. After data analysis, we will find out the data distribution and data types. We will train 4 classification algorithms to predict the output. We will also compare the outputs. Let us get started with the project implementation.

**Superwise Machine learning algorithms and implementation :**

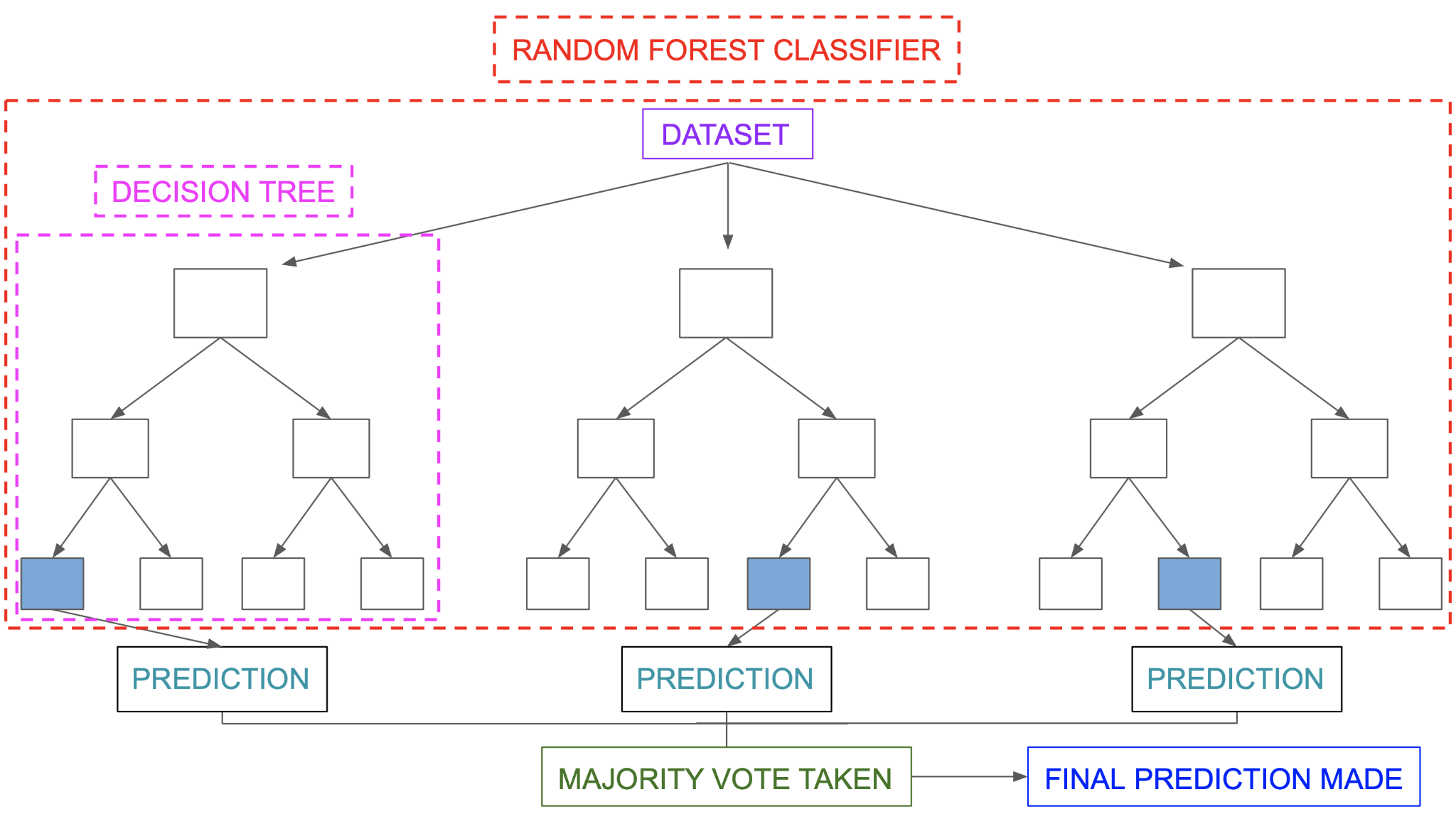
## **Random Forest Classifier:**

A random forest is a supervised machine learning method built from decision tree techniques. This algorithm is used to anticipate behavior and results in a variety of sectors, including banking and e-commerce.

A random forest is a machine learning approach for solving regression and classification issues. It makes use of ensemble learning, which is a technique that combines multiple classifiers to solve complicated problems.

A random forest method is made up of a large number of decision trees. The random forest algorithm’s ‘forest’ is trained via bagging or bootstrap aggregation. Bagging is a meta-algorithm ensemble that increases the accuracy of machine learning algorithms.

The outcome is determined by the (random forest) algorithm based on the predictions of the decision trees. It forecasts by averaging or averaging the output of several trees. The precision of the outcome improves as the number of trees grows.

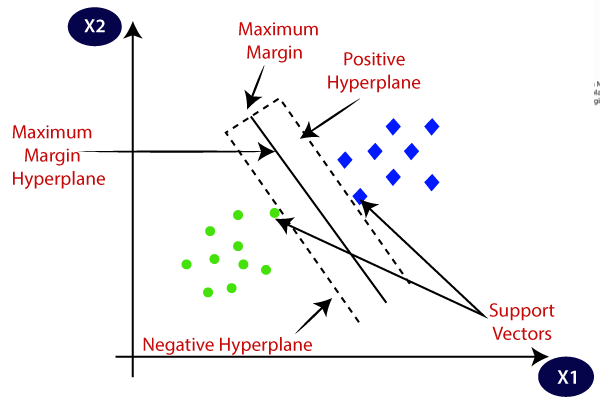


## **SVM Classifier :**

Support Vector Machine, or SVM, is a prominent Supervised Learning technique that is used for both classification and regression issues. However, it is mostly utilized in Machine Learning for Classification purposes.

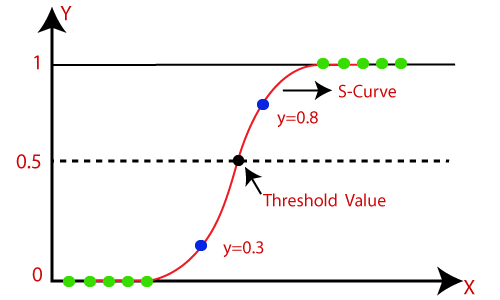
The SVM algorithm’s purpose is to find the optimum line or decision boundary for categorizing n-dimensional space so that we may simply place fresh data points in the proper category in the future. A hyperplane is the optimal choice boundary.

Check [this](https://www.analyticsvidhya.com/blog/2021/06/support-vector-machine-better-understanding/) article for more information on SVM.



**Logistic regression machine learning:**

* 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.
* Logistic Regression is much similar to Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.
* In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).
* The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.
* Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.
* Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:

****

# **Decision Tree Classification Algorithm**

* Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.
* In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
* The decisions or the test are performed on the basis of features of the given dataset.
* *It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.*
* It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
* In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.
* A decision tree simply asks a question, and based on the answer (Yes/No), it further splits the tree into subtrees.
* Below diagram explains the general structure of a decision tree:



**Summary:**

1. From EDA we can see that there are mobile phones in 4 price ranges. The number of elements is almost similar.

2. half the devices have Bluetooth, and half don’t.

3. There is a gradual increase in battery as the price range increases Ram

has continuous increase with price range while moving from Low cost to

Very high cost.

4. costly phones are lighter.

5. RAM, battery power, pixels played more significant role in deciding the

price range of mobile phones.

6. form all the above experiments we can conclude that logistic regression, SVM and Hyperparameter tuning for Random Forest we got the best results

**References:**

● Machine Learning Mastery

● GeeksforGeeks

● Analytics Vidhya Blogs

● Towards Data Science Blogs

● Built in Data Science Blogs

● Scikit- Learn Org

● Investopedia