**Mobile Price Range Prediction**

Nitesh Bhowmick,

Ayush Goyal, M Sameer Ahmed,

Data science trainees

Alma Better, Bangalore

**Abstract:**

The study conducted was on the proposed “Mobile Price Range Prediction”. The main purpose of this study was to create a program that will predict the Mobile Price Range according to features. The researchers used the descriptive type of survey methods where they distributed questionnaires to the respondent of the study as a research instrument for data gathering.

Our experiment can help understand what could be the reason for the classification of such labels by feature selection, data analysis and prediction with Python Programming & ML algorithm taking into account previous trends to determine the correct classification.

**1. Problem Statement**

This data set contains booking information for price range, Battery power, Bluetooth connectivity, RAM, FC (front camera megapixels), PC (Primary camera Megapixels), Mobile weight, and screen size. All personally identifying information has from the data. We will perform exploratory data analysis with python to get insight from the data & using ML algorithm to get the accuracy level of all algorithm .

**2. 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.

I 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.

I have implemented a Mobile Price Prediction using different Machine Learning Algorithms. This project will classify the price range of the mobile price. The price ranges from 0-3. We’ll discuss the price range in the dataset. It's a classification problem. Now I have trained a mobile price classification using 4 ML algorithms. This model classifies the range of the mobile based on the different parameters like from camera, touch screen, cores, battery, clock speed, internal memory, battery capacity, etc. After training the model using 4 algorithms & compared all the models using the accuracy

### 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.

**3. Key features of Mobile Price Range Prediction:-**

**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

**Internal memory** - Internal Memory in Gigabytes

**M\_dep** - Mobile Depth in cm

**Mobile weight** - Weight of mobile phone

**No. of 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).

## **4. 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.

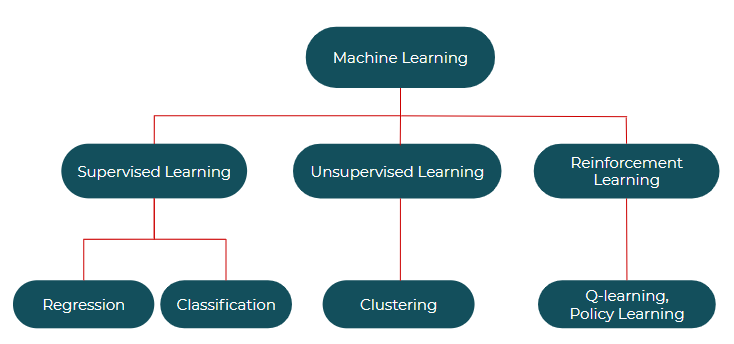
### **5. Machine Learning:**

**Machine Learning:**

Learning is any process by which a system improves performance from experience.

Machine Learning is concerned with computer programs that automatically improve their performance through experience.

**Task in Machine Learning:-**



#### **Supervised learning: -**

#### Supervised learning, as the name indicates, has the presence of a supervisor as a teacher. Basically supervised learning is when we teach or train the machine using data that is well labeled. Which means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that the supervised learning algorithm analyses the training data(set of training examples) and produces a correct outcome from labeled data.

**Types:-**

* Regression
* Logistic Regression
* Classification
* Naive Bayes Classifiers
* K-NN (k nearest neighbors)
* Decision Trees
* Support Vector Machine

**Advantages:-**

* Supervised learning allows collecting data and produces data output from previous experiences.
* Helps to optimize performance criteria with the help of experience.
* Supervised machine learning helps to solve various types of real-world computation problems.

**Disadvantages:-**

* Classifying big data can be challenging.
* Training for supervised learning needs a lot of computation time. So, it requires a lot of time.

#### **Unsupervised learning:-**

Unsupervised learning is the training of a machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here the task of the machine is to group unsorted information according to similarities, patterns, and differences without any prior training of data.

Unlike supervised learning, no teacher is provided that means no training will be given to the machine. Therefore the machine is restricted to find the hidden structure in unlabeled data by itself.

**Types of Unsupervised Learning:-**

**Clustering:-**

1. Exclusive (partitioning)
2. Agglomerative
3. Overlapping
4. Probabilistic

**Clustering Types:-**

1. Hierarchical clustering
2. K-means clustering
3. Principal Component Analysis
4. Singular Value Decomposition
5. Independent Component Analysis

**6. Steps involved:**

* **Exploratory Data Analysis**

 In statistics, exploratory data analysis (*EDA*) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily *EDA* is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

* **Null values Treatment**

Our dataset contains not any null values & Nan values , hence we don’t need to dropped or anything with that dataset. In this dataset there are not any duplicate values.

* **Fitting different models**

For Analysis we tried various types of ML model:

1. **Logistic Regression.**
2. **Random Forest.**
3. **Decision Tree.**

# **Support Vector Machine**

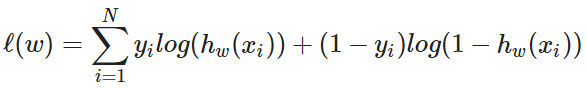
**7. Algorithms:**

1. **Logistic Regression :**

### Logistic regression is a classification algorithm that predicts the probability of an outcome that can only have two values (i.e. a dichotomy). A logistic regression produces a logistic curve, which is limited to values between 0 and 1. Logistic regression models the probability that each input belongs to a particular category.

### Logistic regression is an excellent tool to know for classification problems, which are problems where the output value that we wish to predict only takes on only a small number of discrete values. Here we'll focus on the binary classification problem, where the output can take on only two distinct classes.

Cost Function:-



**Types of Logistic Regression:-**

There are three main types of logistic regression: binary, multinomial and ordinal. They differ in execution and theory. Binary regression deals with two possible values, essentially: yes or no. Multinomial logistic regression deals with three or more values. And ordinal logistic regression deals with three or more classes in a predetermined order.

* **Binary logistic regression**

Binary logistic regression was mentioned earlier in the case of classifying an object as an animal or not an animal—it’s an either/or solution. There are just two possible outcome answers. This concept is typically represented as a 0 or a 1 in coding. Examples include:

* Whether or not to lend to a bank customer (outcomes are yes or no).
* Assessing cancer risk (outcomes are high or low).
* Will a team win tomorrow’s game (outcomes are yes or no).
* **Multinomial logistic regression:-**

Multinomial logistic regression is a model where there are multiple classes that an item can be classified as. There is a set of three or more predefined classes set up prior to running the model. Examples include:

* Classifying texts into what language they come from.
* Predicting whether a student will go to college, trade school or into the workforce.
* Does your cat prefer wet food, dry food or human food?
* **Ordinal logistic regression:-**

Ordinal logistic regression is also a model where there are multiple classes that an item can be classified as; however, in this case an ordering of classes is required. Classes do not need to be proportionate. The distance between each class can vary. Examples include:

* Ranking restaurants on a scale of 0 to 5 stars.
* Predicting the podium results of an Olympic event.
* Assessing a choice of candidates, specifically in places that institute ranked-choice voting.

**Import Logistic Regression:-**

from sklearn.linear\_model import LogisticRegression

.

1. **Random Forest:-**

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of over fitting.

Below are some points that explain why we should use the Random Forest algorithm:

* It takes less training time as compared to other algorithms.
* It predicts output with high accuracy, even for the large dataset it runs efficiently.
* It can also maintain accuracy when a large proportion of data is missing.

## **Advantages of Random Forest**

* Random Forest is capable of performing both Classification and Regression tasks.
* It is capable of handling large datasets with high dimensionality.
* It enhances the accuracy of the model and prevents the over fitting issue.

## **Disadvantages of Random Forest**

* Although random forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks.

**Import Random Forest Classifier:-**

from sklearn.ensemble import RandomForestClassifier

1. **Decision Tree:-**

# Decision Tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

# **Terminology:-**

* **Root Node**: It represents entire population or sample and this further gets divided into two sets.
* **Splitting**: It is a process of dividing a node into two sub-nodes.
* **Decision Node:** When a sub-node splits into further sub-nodes, then it is called decision node.
* **Leaf/ Terminal Node:** Nodes do not split is called leaf or terminal node.
* **Pruning:** When we remove sub-nodes of a decision node, this process is called pruning. You can say opposite process of splitting.
* **Branch / Sub-Tree:** A sub section of entire tree is called branch or sub-tree.
* **Parent and Child Node:** A node, which is divided into sub-nodes is called parent node of sub-nodes where as sub-nodes are the child of parent node.

**Construction of Decision Tree:**

A tree can be “learned” by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions. The construction of a decision tree classifier does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high-dimensional data. In general decision tree classifier has good accuracy. Decision tree induction is a typical inductive approach to learn knowledge on classification.

**Decision Tree Representation:**

Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, and then moving down the tree branch corresponding to the value of the attribute as shown in the above figure. This process is then repeated for the sub tree rooted at the new node.

**Strengths and Weaknesses of the Decision Tree approach:-**

**The strengths of decision tree methods are:**

* Decision trees are able to generate understandable rules.
* Decision trees perform classification without requiring much computation.
* Decision trees are able to handle both continuous and categorical variables.
* Decision trees provide a clear indication of which fields are most important for prediction or classification.

**The weaknesses of decision tree methods are:**

* Decision trees are less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute.
* Decision trees are prone to errors in classification problems with many classes and a relatively small number of training examples.
* Decision tree can be computationally expensive to train. The process of growing a decision tree is computationally expensive. At each node, each candidate splitting field must be sorted before its best split can be found. In some algorithms, combinations of fields are used and a search must be made for optimal combining weights. Pruning algorithms can also be expensive sincemany candidate sub-trees must be formed and compared.

**Import Decision Tree Classifier:-**

from sklearn.tree import DecisionTreeClassifier

# **Support Vector Machine:-**

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane

**Advantages of support vector machines are:**

* Effective in high dimensional spaces.
* Still effective in cases where number of dimensions is greater than the number of samples.
* Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
* Versatile: different [Kernel functions](https://scikit-learn.org/stable/modules/svm.html#svm-kernels) can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

**Disadvantages of support vector machines are:**

* If the number of features is much greater than the number of samples, avoid over-fitting in choosing [Kernel functions](https://scikit-learn.org/stable/modules/svm.html#svm-kernels) and regularization term is crucial.
* SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation

## **Types of SVM:-**

SVM can be of two types:

* **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.
* **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

## **Hyperplane and Support Vectors in the SVM algorithm:**

**Hyperplane:** There can be multiple lines/decision boundaries to segregate the classes in n-dimensional space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the hyperplane of SVM.

The dimensions of the hyperplane depend on the features present in the dataset, which means if there are 2 features (as shown in image), then hyperplane will be a straight line. And if there are 3 features, then hyperplane will be a 2-dimension plane.

We always create a hyperplane that has a maximum margin, which means the maximum distance between the data points.

**Support Vectors:**

The data points or vectors that are the closest to the hyperplane and which affect the position of the hyperplane are termed as Support Vector. Since these vectors support the hyperplane, hence called a Support vector.

# **Import Support Vector Classifier**

from sklearn import svm

from sklearn.svm import SVC

**8. Conclusion:**

* From EDA we can see that there are mobile phones in 4 price ranges. The number of elements is almost similar.
* Half of devices have Bluetooth connectivity & another half of devices don’t have Bluetooth connectivity.
* 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
* costly phones are lighter
* RAM, battery power, pixels played and Connectivity feature 'three\_g’ & 'four\_g' more significant role in deciding the price range of mobile phone.
* Form all the above experiments we can conclude that logistic regression and SVM with using hyper parameters we got the best results.

**References-**

1. Geeks for Geeks
2. Analytics Vidhya
3. Nvidia
4. scikit-learn