MOBILE PRICE PREDICTION

Rupal Sonje (18IT1086)

Soham Salkar (18IT1050)

Riteshkumar Singh (18IT1068)

Information Technology Department

Ramrao Adik Institute of Technology

Nerul, Navi Mumbai.

***Abstract*— In this Modern Era, Smartphones are an integral part of the lives of human beings. When a smartphone is purchased, many factors like the Display, Processor, Memory, Camera, Thickness, Battery, Connectivity, and others are taken into account. One factor that people do not consider is whether the product is worth the cost. To solve the problem we will develop a model that will predict the approximate price of the new smartphone on the basis of given data. Battery power, clock\_speed, Front Camera, Mobile Depth, Number of cores, Resolution Height, Resolution Width, RAM, Screen Height, Screen Width, Bluetooth, 4G/3G Supported, Touch Screen, etc.**

***Keywords—Model; Smartphone;PricePrediction;***

1. **INTRODUCTION**

Price is the most effective attribute of marketing and business. The very first question of costumer is about the price of items. All the costumers are first worried and thinks “If he would be able to purchase something with given specifications or not”

Artificial Intelligence-which makes machine capable to answer the questions intelligentlynow a days is very vast engineering field. Machine learning provides us best techniques for artificial intelligence like classification, regression, supervised learning and unsupervised learning and many more. Different tools are available for machine learning tasks like MATLAB, Python, Cygwin, WEKA etc. We can use any of classifiers like, Linear Regression, KNN and many more. Different type of feature selection algorithms are available to select only best features and minimize dataset. This will reduce computational complexity of the problem. As this is optimization problem so many optimization techniques are also used to reduce dimensionality of the dataset

Mobile now a days is one of the most selling and purchasing device. Every day new mobiles with new version and more features are launched. Hundreds and thousands of mobile are sold and purchased on daily basis. So here the mobile price class prediction is a case study for the given type of problem i.e. finding optimal product. The same work can be done to estimate real price of all products like cars, bikes, generators, motors, food items, medicine etc.

Many features are very important to be considered to estimate price of mobile. For example Processor of the mobile. Battery timing is also very important in today’s busy schedule of human being. Size and thickness of the mobile are also important decision factors. Internal memory, Camera pixels, and video quality must be under consideration. Internet browsing is also one of the most important constraints in this technological era of 21st century. And so is the list of many

features based upon those, mobile price is decided. So we will use many of above mentioned features to classify whether the mobile would be very economical, economical, and expensive or very\_ expensive.

The structure of the paper is as follows. Next section is review of previous work.3 rd Section contains Methodology and Experimental procedure. Section 4 is the summary of the results. Comparative study is done in section 5. After that paper is concluded in section 6. Outcomes of the work are discussed in section 7. At last in 8th section some suggestions about future work are given.

1. **RELATED WORK**

Using previous data to predict price of available and new launching product is an interesting research background for machine learning researchers. Sameerchand-Pudaruth [1] predict the prices of second hand cars in Mauritius. He implemented many techniques like Multiple linear regression, k-nearest neighbors(KNN), Decision Tree, and Naïve Bayes to predict the prices. Sameerchand-Pudaruth got Comparable results from all these techniques. During research it was found that most popular algorithms i.e Decision Tree and Naïve Bayes are unable to handle, classify and predict Numerical values. Number of instances for his research was only 97(47 Toyota+38 Nissan+12 Honda). Due to less number of instances used, very poor prediction accuracies were recorded[1].

Shonda Kuiper[2] has also worked in the same field. Kuiper used multivariate regression model to predict price of 2005 General Motor cars. He collected the data from available online source www.pakwheels.com. The main part of this research work is “Introduction of suitable variable selection techniques, which helped to find that which variables are more suitable and relevant for inclusion in model. This (His research) helps students and future researchers in many fields to understand the conditions under which studies should be conducted and gives them the knowledge to discern when appropriate techniques should be used[2].

Support Vector Machine(SVM) concept is used by one another researcher Mariana Listiani [3] for the same work. Listiani predicted prices of leased cars using above mentioned technique. It was found in this research that SVM technique is far more better and accurate for price prediction as compared to other like multiple linear regression when a very large data set is available. The researcher also showed that SVM also

handles high dimensional data better and avoids both the under-fitting and over-fitting issues. To find important features for SVM Listiani used Genetic Algorithm . However, the technique failed to show in terms of variance and mean standard deviation why SVM is better than simple multiple regression[3].

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Description** | **Possible Values** |
| BATTERY POWER | Mobile’s battery | { } |
| BLUETOOTH | Does mobile have Bluetooth | {0,1 } |
| CLOCK SPEED MHz | What is clock speed of mobile in megahertz | { } |
| DUAL SIM | Does mobile support dual sim | {0,1} |
| FRONT CAMERA | What is megapixel of front camera of mobile | {} |
| 4G ENABLED | Does mobile support 4g or not | {0,1} |
| INTERNAL MEMORY(GB) | Internal memory of mobile in gb | { } |
| DEPTH(CM) | Depth of a mobile | { } |
| MOBILE WEIGHT(GM) | What is mobile weight in grams |  |
| NO OF CORES | No of cores present in mobile |  |

Neural Networks (NN) are more better in estimating price of house, this was concluded in the research of Limsombunchai[4]. By comparing with hedonic method his method was more accurate.Operation of both the methods are same, but in NN the model is trained first and then tested for prediction. Using both the methods NN produced higher R-sq and smaller root mean square error (RMSE), while hedonic produced lower values. This research was limited because the actual house price were missing and only estimated prices were used for the research work[4].

K Noor and Saddaqat J [5] also worked to predict the price of Vehicles using different techniques. The researchers achieved highest accuracy using multiple linear regression. This paper proposes a system where price is dependent variable which is predicted, and this price is derived from factors like vehicle’s model, make, city, version, color, mileage, alloy rims and power steering[5].

1. **DATA MINING PROCESS**

Data

Collection

Preprocess

ing

Apply

Algorithm

Accuracy

of Result

1. *Dataset*

The features of mobiles are collected from [https://www.kaggle.com](https://www.kaggle.com/)

TABLE I. ATTRIBUTES DESCRIPTION AND POSSIBLE VALUES

|  |  |  |
| --- | --- | --- |
| PRIMARY CAMERA(MP) | What is megapixel of primary camera of mobile |  |
| PIXEL HEIGHT | Pixel height of mobile |  |
| PIXEL WIDTH | Pixel width of mobile |  |
| RAM (MB) | Ram of mobile |  |
| SCREEN HEIGHT (CM) | Screen height of mobile in cm |  |
| SCREEN WIDTH(CM) | Screen width of mobile in cm |  |
| TALKTIME (HR) | Talk time of a mobile |  |
| 3G ENABLED | Is 3g enabled in mobile | {0,1} |
| TOUCH SCREEN | Is mobile touchscreen | {0,1} |
| WIFI | Is Wi-Fi present in mobile | {0,1} |

* touch\_screen:touch screen supported or not
* wifi: wifi supported 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).

*B. Data Exploration*

In order to understand the dataset in hand, it must be explored in a statistical manner, as well as, visualize it using graphical plots and diagrams. This step in data mining is essential because it allows the researchers as well as the readers to understand the data before jumping into applying more complex data mining tasks and algorithms.

Table 2 shows the ranges of the data in the dataset according to their attributes, ordered from highest to lowest.

TABLE II. RANGES OF DATA IN THE DATASE

Following is a more detailed description about some attributes mentioned in Table 1:

* battery\_power:Total energy a battery can store in one time measured in mAh
* clock\_speed:The speed at which microprocessor executes instructions
* fc Front Camera:megapixels
* int\_memory:Internal Memory in Gigabytes
* m\_dep:Mobile Depth in cm
* mobile\_wt:Weight of the mobile phone
* n\_cores:Number of cores of a processor
* pc Primary Camera:megapixels
* px\_height:Pixel Resolution Height
* px\_width:Pixel Resolution Width
* ram:Random Access Memory in MegaBytes
* sc\_h:Screen Height of mobile in cm
* sc\_w:Screen Width of mobile in cm
* talk\_time:The longest time that a single battery charge will last price\_range
* blue: has Bluetooth or not
* four\_g: 4g supported or not
* three\_g: 3g supported or not

|  |  |
| --- | --- |
| **Attribute** | **Range** |
| BATTERY POWER | {500,5000} |
| CLOCK SPEED MHz | (50,500} |
| FRONT CAMERA | {3,15} |
| INTERNAL MEMORY(GB) | {1,8} |
| DEPTH(CM) | {2,4} |
| MOBILE WEIGHT(GM) | {80,600} |
| NO OF CORES | {1,7} |
| PRIMARY CAMERA(MP) | {1,15} |
| PIXEL HEIGHT | {800.1800} |
| PIXEL WIDTH | {200.1000} |
| RAM (MB) | {1000.6000} |
| SCREEN HEIGHT (CM) | {800.1800} |
| SCREEN WIDTH(CM) | {200.1000} |

Analysis of parameters:

Fig1: Descriptive Statistics of parameters



Fig2: 3G/4G supported mobiles

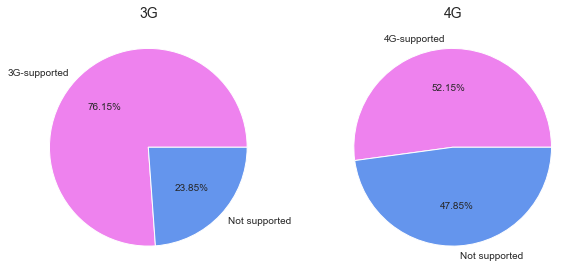


Fig3: cores in mobile

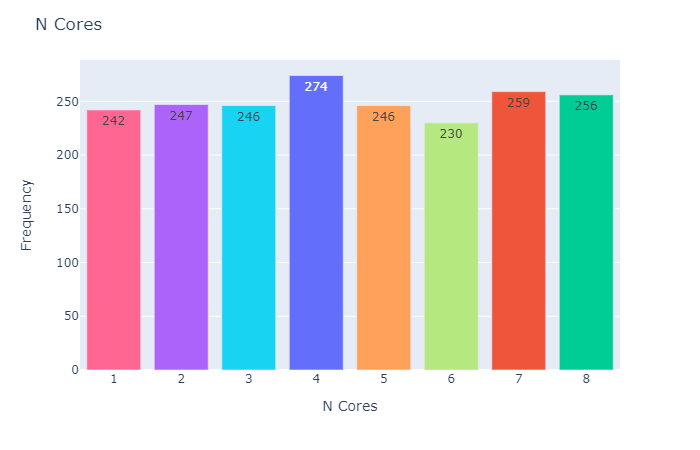
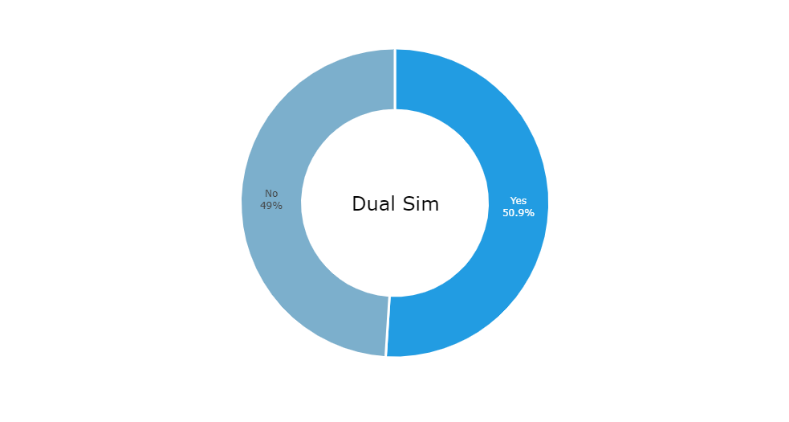
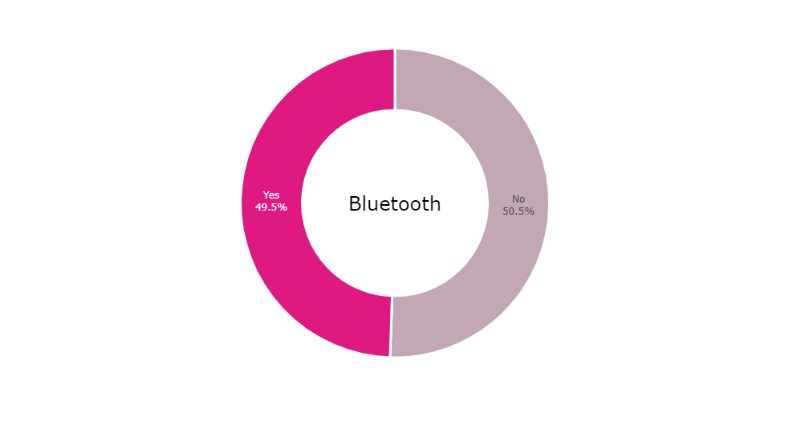


Fig4: Dual Sim supported mobiles

 Fig5,6: Bluetooth supported mobiles



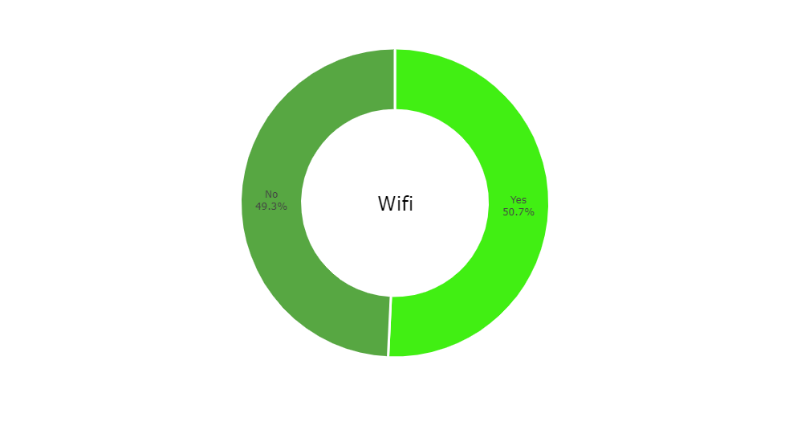


Fig7: RAM in mobile phones

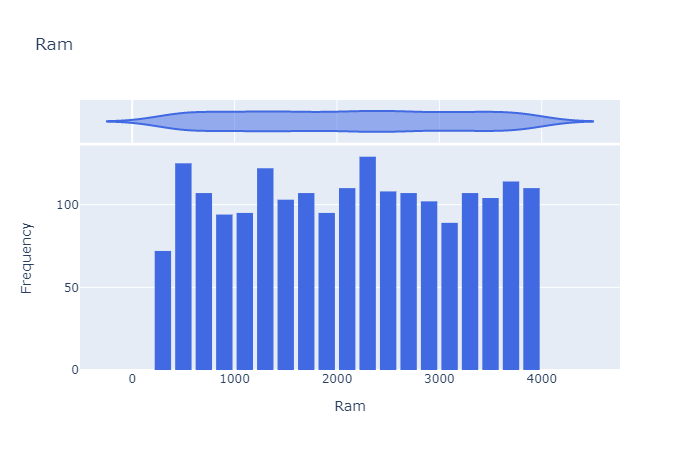


Fig8: Battery Power analysis

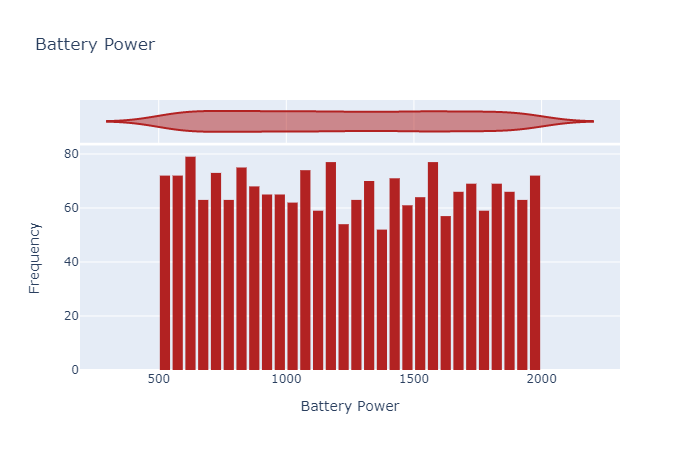


Fig9: Internal Memory analysis

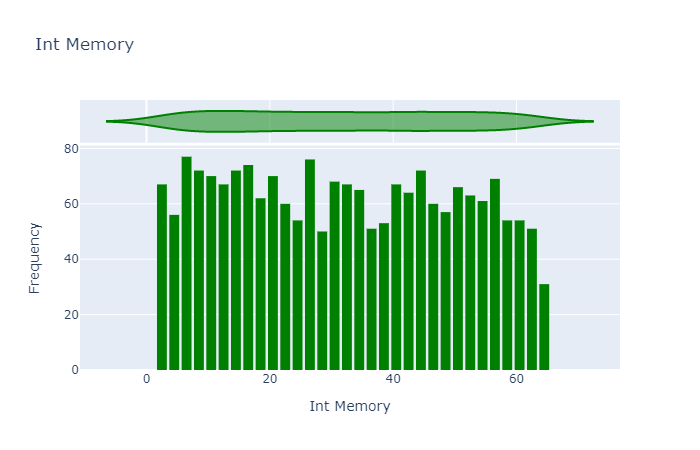
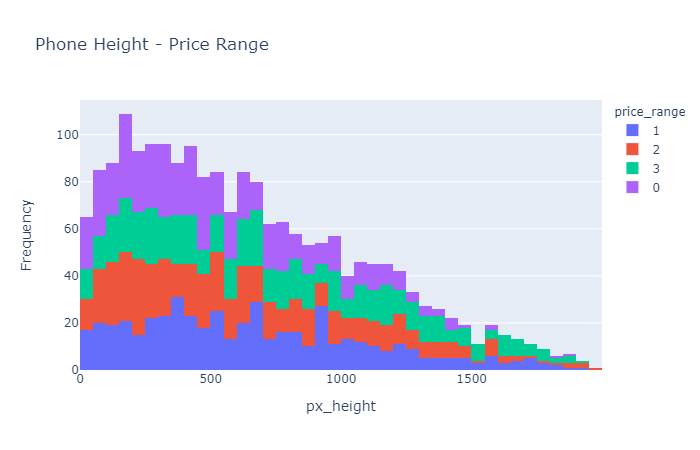


Fig10: Phone Height Power analysis

 Fig11: Phone Width analysis

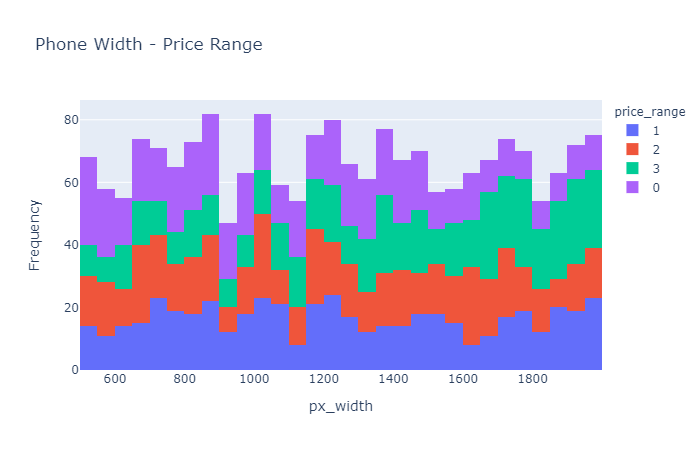


Fig12: Ram effect on Mobile Price

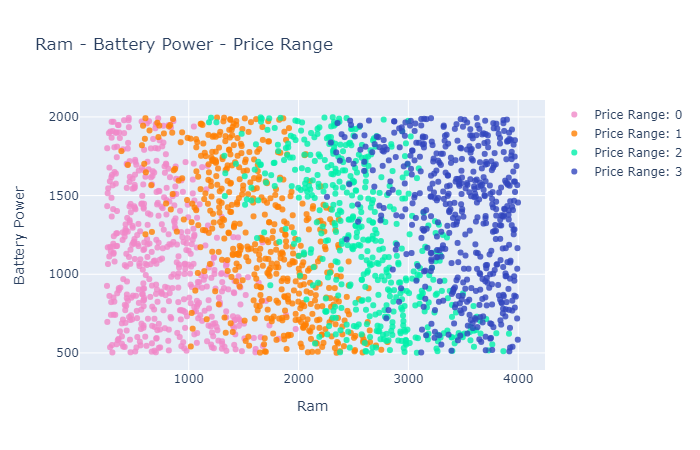
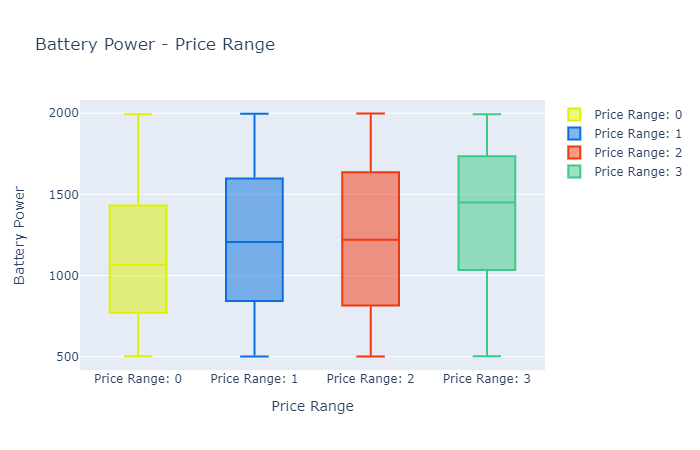


Fig13: Battery Power effect on Mobile Price



D. DATA MINING IMPLEMENTATION

There are multiple well known techniques available for data mining and knowledge discovery in databases (KDD), such as Regression, Classification, Clustering, Association Rule Learning, Artificial Intelligence, etc.

Regression is one of the mostly used and studied data mining technique. Researchers use and study regression because it is simple and easy to use. In detail, in data mining, Regression is a technique for predicting a continuous outcome variable based on the value of one or more variables previously learned classes from a training dataset, where the classes of the objects are known. There are multiple regression techniques available in data mining, such as, Linear regression, Logistic Regression, Random Forest Regression, Bayesian Linear Regression.

In this study, multiple regression techniques were used in the data mining process for predicting the price of a mobile phone. This approach was used because it can provide a broader look and understanding of the all parameters All data mining implementation and processing in this study was done using jupyter Notebook.

1. Logistic Regression

Logistic Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It’s used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog). There are two main types: Simple regression and multivariable regression.

2. K-Nearest Neighbor

KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbourhood. The size of the neighbourhood needs to be set by the analyst or can be chosen using cross-validation (we will see this later) to select the size that minimises the mean-squared error.x

The KNN algorithm uses ‘feature similarity’ to predict the values of any new data points. This means that the new point is assigned a value based on how closely it resembles the points in the training set. The steps involved in KNN method are: calculating the distance between new point and each training point, select the k and make predictions.

3. Random Forest Regression

Random Forest Regression 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.

A Random Forest Regression model is powerful and accurate. It usually performs great on many problems, including features with non-linear relationships. To get a better understanding of the Random Forest algorithm, let’s walk through the steps:

1.Pick at random k data points from the training set.

2.Build a decision tree associated to these k data points. 3.Choose the number N of trees you want to build and repeat steps 1 and 2.

4.For a new data point, make each one of your N-tree trees predict the value of y for the data point in question and assign the new data point to the average across all of the predicted y values.

4. Support Vector Clustering:

The objective of [clustering](http://www.scholarpedia.org/w/index.php?title=Data_Clustering&action=edit&redlink=1) is to partition a data set into groups according to some criterion in an attempt to organize data into a more meaningful form. There are many ways of achieving this goal. Clustering may proceed according to some parametric model or by grouping points according to some distance or similarity measure as in [hierarchical clustering](http://www.scholarpedia.org/w/index.php?title=Hierarchical_Cluster_Analysis&action=edit&redlink=1). A natural way to put cluster boundaries is in regions in data space where there is little data, i.e. in "valleys" in the probability distribution of the data. This is the path taken in support vector clustering (SVC), which is based on the [support vector](http://www.scholarpedia.org/w/index.php?title=Support_Vector_Machine&action=edit&redlink=1) approach (see [Ben-Hur et al., 2001](http://www.scholarpedia.org/article/Support_vector_clustering#benhur2001)).

In SVC data points are mapped from data space to a high dimensional feature space using a kernel function. In the kernel's feature space the [algorithm](http://www.scholarpedia.org/article/Algorithm) searches for the smallest sphere that encloses the image of the data using the [Support Vector Domain Description](http://www.scholarpedia.org/w/index.php?title=Support_Vector_Domain_Description&action=edit&redlink=1) algorithm. This sphere, when mapped back to data space, forms a set of contours which enclose the data points. Those contours are then interpreted as cluster boundaries, and points enclosed by each contour are associated by SVC to the same cluster.

5. Decision Tree:

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.*

# Analysis and Summary

# FIG : CO RELATION MATRIX

# C:\Users\Admin\Downloads\4.png

# Fig: Algorithm AnalysisC:\Users\Admin\Downloads\5.png

# Fig: Accuracy of each Algorithm

# C:\Users\Admin\Downloads\6.png

In this section, multiple algorithm techniques and their performances and accuracies were tested and validated. As a final analysis, it was obviously noticed that some algorithms worked better with the dataset than others, in detail, Logistic Regression Algorithm had the best accuracy of 97.01%, which was significantly more than the expected (default model) accuracy, SVC was next with 94.07% respectively, and the least accurate was KNN with 64.80%.

**CONCLUSION:**

This work can be concluded with the comparable results of both Feature selection algorithms and classifier. This combination has achieved maximum accuracy and selected minimum but most appropriate features. It is important to note that in Forward selection by adding irrelevant or redundant features to the data set decreases the efficiency of both classifiers. While in backward selection if we remove any important feature from the data set, its efficiency decreases. The main reason of low accuracy rate is low number of instances in the data set. One more thing should also be considered while working that converting a regression problem into classification problem introduces more error

**OUTCOMES OF THE WORK**

Cost prediction is the very important factor of marketing and business. To predict the cost same procedure can be performed for all types of products for example Cars, Foods, Medicine, Laptops etc.

Best marketing strategy is to find optimal product (with minimum cost and maximum specifications). So products can be compared in terms of their specifications, cost, manufacturing company etc.

By specifying economic range a good product can be suggested to a costumer.

**FUTURE WORK EXTENSION**

More sophisticated artificial intelligence techniques can be used to maximized the accuracy and predict the accurate price of the products.

Software or Mobile app can be developed that will predict the market price of any new launched product.

To achieve maximum accuracy and predict more accurate, more and more instances should be added to the data set. And selecting more appropriate features can also increase the accuracy. So data set should be large and more appropriate features should be selected to achieve higher accuracy.

**ACKNOWLEDGMENT**

We would take this opportunity to express our sincere gratitude to Ramrao Adik Institute of Technology, Nerul for giving us the opportunity to explore more on the topic “Student Performance in Exam” as a part of our Data mining mini project.. We are grateful to our respected Principal Sir Dr. Mukesh. D. Patil, for his support and guidance. We would like to thank Dr. Ashish Jadhav sir, HOD Information Technology for his support. We would like to profusely thank Mrs. Jyoti Deone ma’am without whom this project would have been a distant reality. Thanks and appreciation to our classmates for their constant encouragement and support.

**REFERENCES**

1. Sameer Chand Pudaruth. “Predicting the Price of Used Cars using Machine Learning Techniques”, International Journal of Information & Computation Technology. ISSN 0974-2239 Volume 4, Number 7 (2014), pp. 753764

2. Shonda Kuiper, “Introduction to Multiple Regression: How Much Is Your Car Worth? ” , Journal of Statistics Education · November 2008

3. Mariana Listiani, 2009. “Support Vector Regression Analysis for Price Prediction in a Car Leasing Application”. Master Thesis. Hamburg University of Technology

4. https://www.kaggle.com/vikramb/mobileprice-prediction

5. file:///C:/Users/Lenovo/Downloads/asim2018-ijca-9165551%20(1).pdfMuhammad Asim UET Lahore Pakistan, Zafar Khan UET Lahore Pakistan

6. Introduction to dimensionality reduction, A computer science portal for Geeks. https://www.geeksforgeeks.org/dimensionalityreduction