# ABSTRACT

To predict “If the mobile with given features will be Economical or Expensive” is the main motive of this research work. Real Dataset is collected from website www.kaggle.com. 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.

# 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”. So to estimate price at home is the basic purpose of the work. This paper is only the first step toward the above mentioned destination.

Artificial Intelligence-which makes machine capable to answer the questions intelligently- now 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 Decision tree , Naïve Bayes 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 mobiles 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 todays 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, expensive or very\_ expensive.

**Methodology:**

The experiment is performed using Jupyter notebook. The main steps of machine learning are as follows:

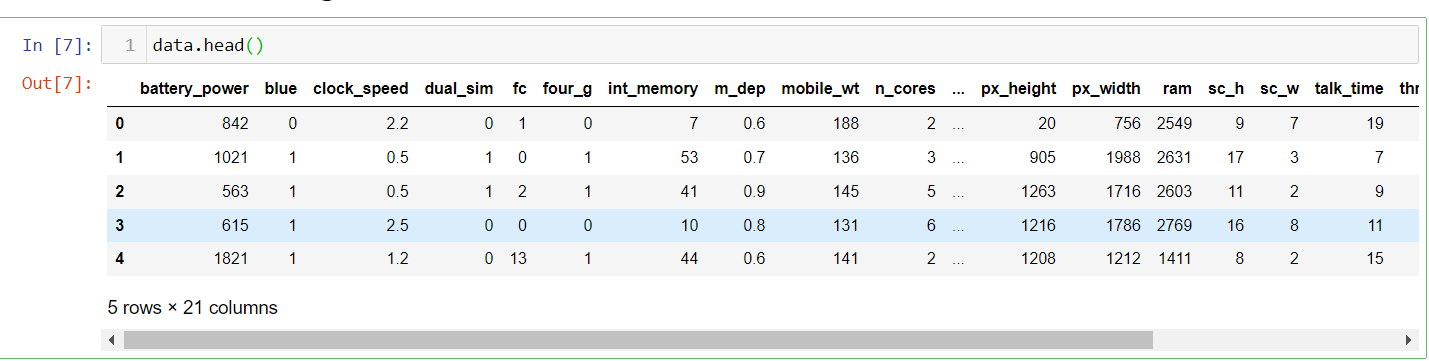
1:Data Collection

2: Dimensionality Reduction

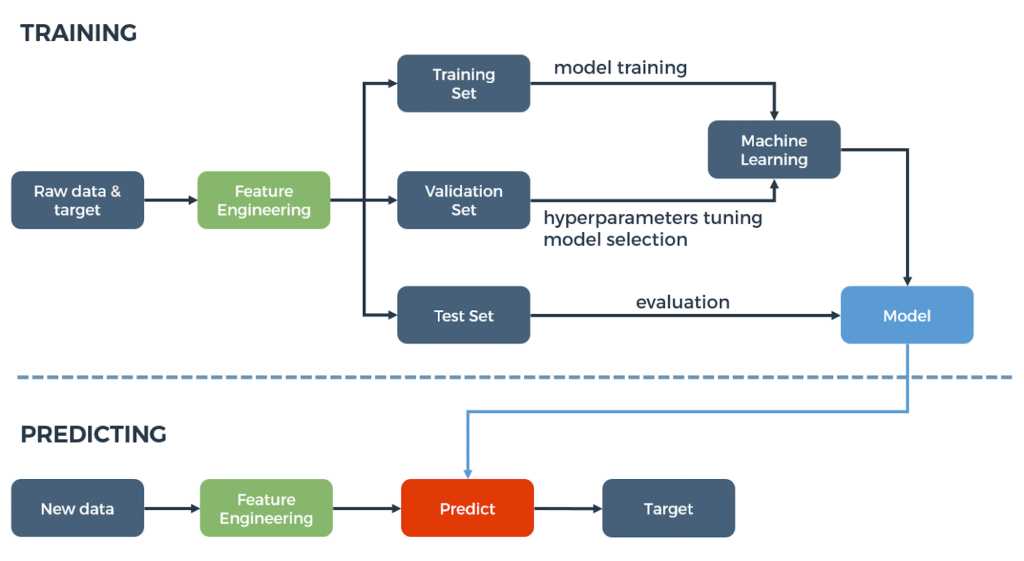
3: Classification

1: Data Collection:

The dataset with shape(2000,21) is selected from a Kaggle website.



The below passages describe about the methodology used in the real estate house price predictions and the architecture flow diagram is given



**2: Dimensionality Reduction:**

Dimensionality reduction is the process of reducing the number of random variables(Features) under consideration, by obtaining a set of principal variables. The higher the number of features, the harder it gets to visualize the training set and then work on it. Sometimes, most of these features are correlated, and hence redundant. This is where dimensionality reduction algorithms come into play.

Two types of Dimensionality reduction algorithms are there ie

1: Feature selection,

2: Feature extraction.

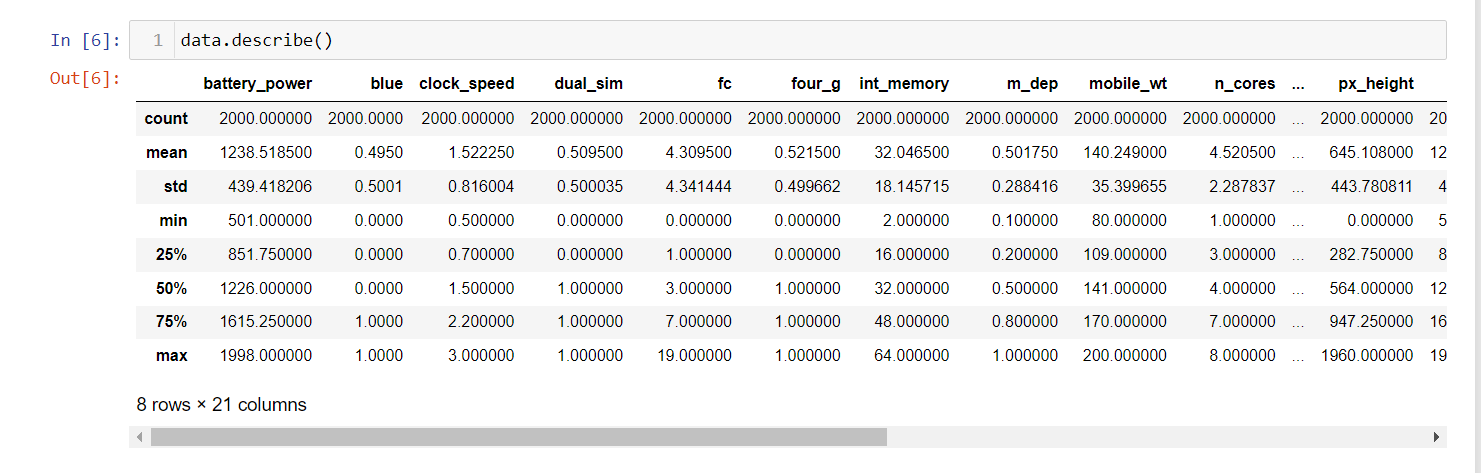
**Feature Selection**

In feature selection we are interested in finding k of the d dimensions that give us the most information, and we discard the other (d − k) dimensions.

**Feature Extraction**

In feature extraction we are interested in finding a new set of k dimensions that are combinations of the original d dimensions for example Principal Component Analysis. Here feature selection algorithms are used. There are two approaches: Forward selection and backward selection. Forward Selection In forward selection, we start with no variables and add them one by one, at each step adding the one that decreases the error the most, until any further addition does not decrease the error (or decreases it only slightly). Backward Selection In backward selection we start with all variables and remove them one by one, at each step removing the one that decreases the error the most (or increases it only slightly), until any further removal increases the error significantly.

**Dataset Review:**

A dataset in machine learning is, quite simply, a collection of data pieces that can be treated by a computer as a single unit for analytic and prediction purposes. I have used one dataset in this paper where various existing machine learning algorithms are applied to the datasets for predicting prices. The dataset is from the Kaggle which concerns price prediction. This dataset compromise of battery power, clock\_speed,mobile\_wt,n\_cores etc. As this paper uses machine learning for price prediction, attribute variables are used to predict the label/price. The following table shows the set of attribute variables to develop the prediction model. 

**Dataset preprocessing:**

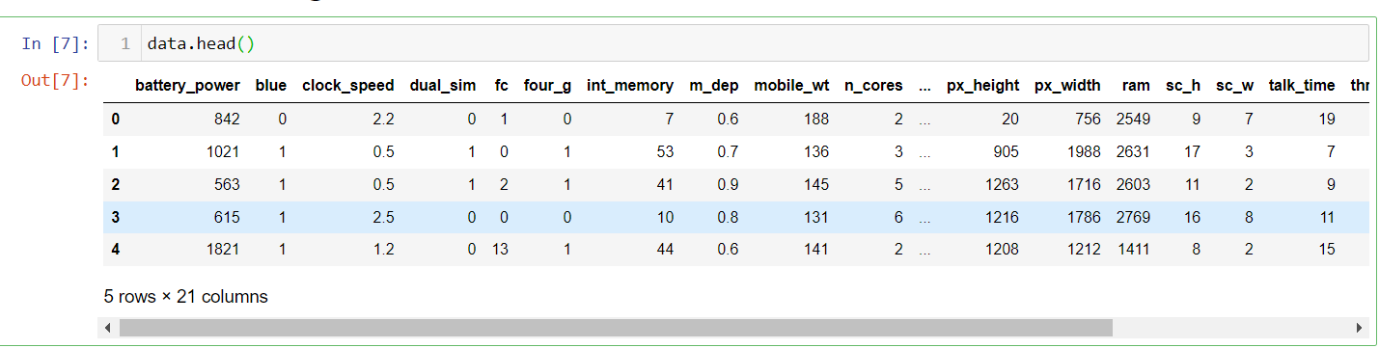
Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data pre-processing task.

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data pre-processing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

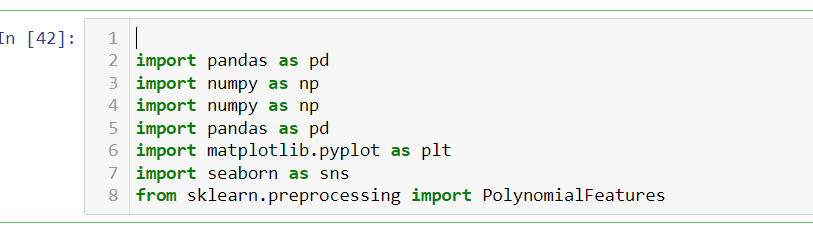
It involves the following steps:

* Getting the dataset
* Importing libraries
* Importing datasets
* Finding Missing Data
* Splitting dataset into training and test set
* Feature scaling

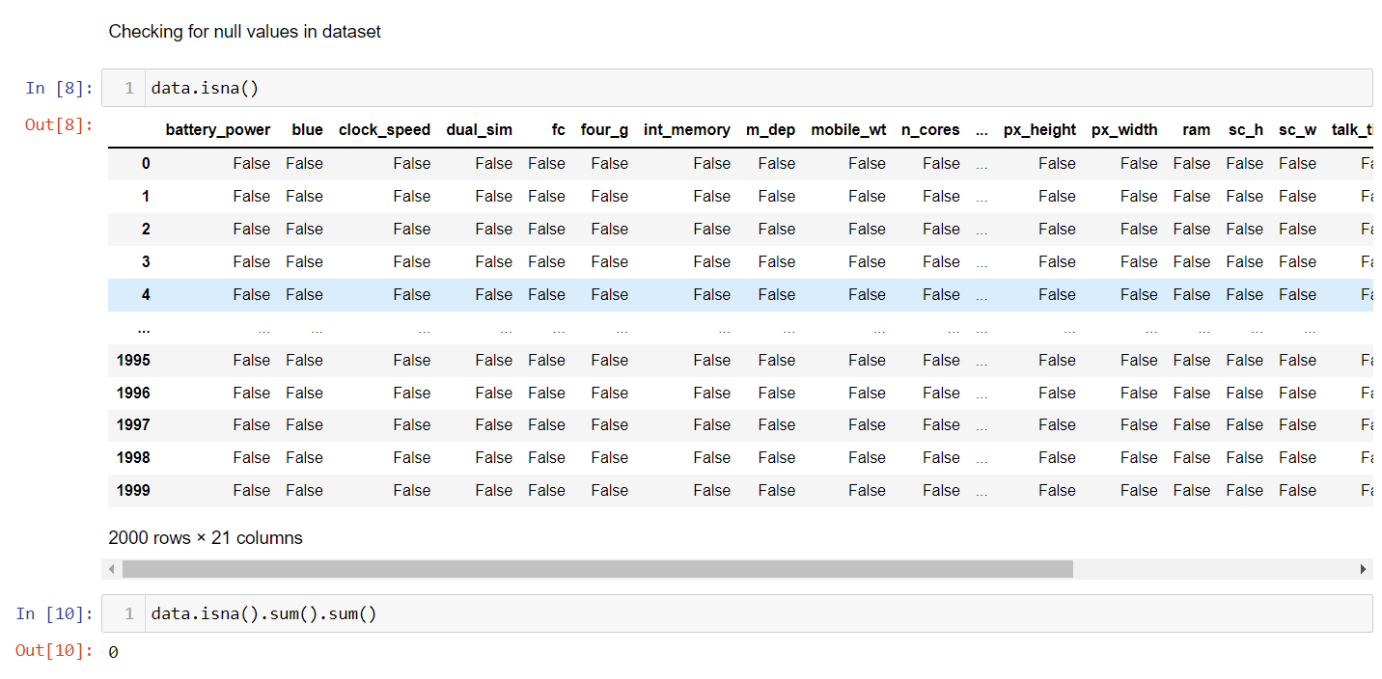
Getting Dataset:



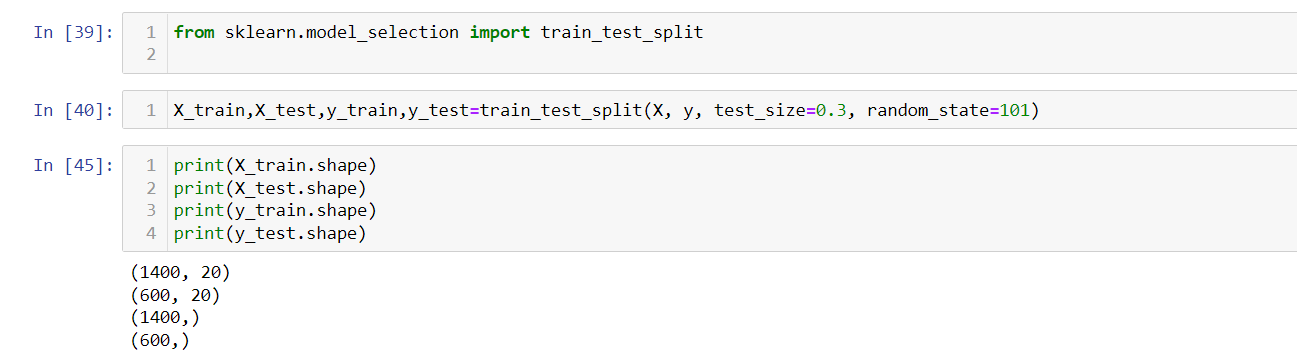
Importing libraries:

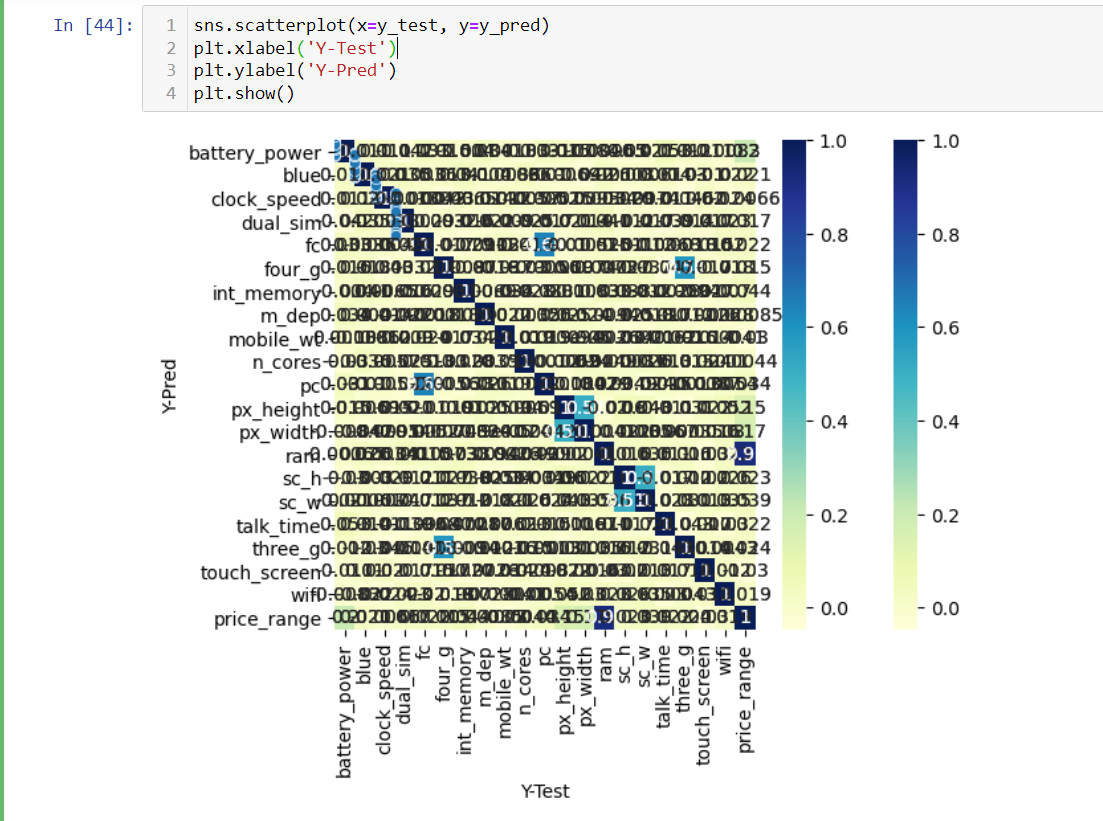


Finding Missing Data:



Splitting the dataset into training and testing:



FeatureScaling::

**Algorithms used:**

1. **Linear Regression**
2. **Decision Tree Regression**
3. **Random Forest Regression**
4. **Support Vector Regression**
5. **Extreme Gradient Boost Regression**

**Linear Regression:**

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

**Decision Tree Regression:**

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**

A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as **bagging**. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

**Support Vector Regression:**

Supervised Machine Learning Models with associated learning algorithms that analyze data for classification and regression analysis are known as Support Vector Regression. SVR is built based on the concept of Support Vector Machine or SVM. It is one among the popular Machine Learning models that can be used in classification problems or assigning classes when the data is not linearly separable.

**The assembled model that shows all the MSE\_score and R2\_score as follows:**

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No** | **Algorithm** | **MSE\_score** | **R2\_score** |
| **1** | Linear Regression | 0.109440 | 0.087843 |
| **2** | Support Vector Regression | 0.068615 | 0.949191 |
| **3** | Decision Tree Regression | 0.167500 | 0.875967 |
| **5** | Random Forest Regression | 0.090485 | 0.932996 |
| **5** | XGB Regression | 0.087843 | 0.934952 |

**B**y observing the above model I am choosing support vector regression as a best model because it having less MSE\_score and high R\_2 score.

**Mean square error , R2 score of accuracy of the validation, test, train datasets:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Datasets** | **Mean square error** | **R2 score** | **accuracy** |
| **Test dataset** | **0.2619** | **0.9491** | **0.9428** |
| **Train dataset** | **-** | **-** | **0.9416** |
| **Validation dataset** | **2.560** | **-1.110** | **0.9413** |

**Conclusion:**

This work can be concluded with the comparable results all the algorithms. As we seen that the support vector regression is giving the best results comparing to the other algorithms. This model is used to predict the prices of the mobiles.

**References:**

1. Medium.com
2. Kaggle.com
3. <https://www.geeksforgeeks.org/machine-learning>
4. [ML - Wikipedia](https://en.wikipedia.org/wiki/ML)
5. towardsdatascience.com

Git-hub: