# MACHINE LEARNING

## (Predicting Price Ranges of Mobile Phones)

Summer Internship Report Submitted in partial fulfillment of the requirement for undergraduate degree of

**Bachelor of Technology In**

**Computer Science Engineering**

By

**BEACHANI SHARATH KUMAR**

**221710309009**

*Under the Guidance of*

## Mr.

Assistant Professor



Department Of Computer Science Engineering GITAM School of Technology

GITAM (Deemed to be University) Hyderabad-502329 July 2020

i

## DECLARATION

I submit this industrial training work entitled “PREDICTING PRICE RANGES OF MOBILE PHONES” to GITAM (Deemed To Be University), Hyderabad in partial fulfillment of the requirements for the award of the degree of “Bachelor of Technology” in “Computer Science Engineering”. I declare that it was carried out independently by me under the guidance of Mr. , Asst. Professor, GITAM (Deemed To Be University), Hyderabad, India.

The results embodied in this report have not been submitted to any other University or Institute for the award of any degree or diploma.

Place: HYDERABAD BEACHANI SHARATH KUMAR

Date:14-07-2020 Roll No. 221710309009



ii)

GITAM (DEEMED TO BE UNIVERSITY

Hyderabad-502329,India Dated:

CERTIFICATE

This is to certify that the Industrial Training Report entitled “PREDICTING PRICE RANGES OF MOBILE PHONES” is being submitted by BEACHANI SHARATH KUMAR (221710309009) in partial fulfillment of the requirement for the award of Bachelor of Technology in Computer Science Engineering at GITAM (Deemed To Be University), Hyderabad during the academic year 2020-21

It is faithful record work carried out by him at the Computer Science Engineering Department, GITAM University Hyderabad Campus under my guidance and supervision.

Mr. Dr.S.Phani Kumar

Assistant Professor Professor and HOD

Department of CSE Department of CSE

iv

## ACKNOWLEDGEMENT

Apart from my effort, the success of this internship largely depends on the encouragement and guidance of many others. I take this opportunity to express my gratitude to the people who have helped me in the successful competition of this internship.

I would like to thank Dr. N. Siva Prasad, Pro Vice Chancellor, GITAM Hyderabad and Dr.

N .Seetharamaiah, Principal, GITAM Hyderabad

I would like to thank respected Dr. S.PhaniKumar, Head of the Department of Computer Science Engineering for giving me such a wonderful opportunity to expand my knowledge for my own branch and giving me guidelines to present an internship report. It helped me a lot to realize of what we study for

I would like to thank the respected faculties Mr. who helped me to make this internship a successful accomplishment.

I would also like to thank my friends who helped me to make my work more organized and well-stacked till the end

BEACHANI SHARATHKUMAR

Roll.No. 221710305036

v

## ABSTRACT

Machine learning algorithms are used to predict the values from the data set by splitting the data set in to train and test and building Machine learning algorithms models of higher accuracy to predict the values is the primary task to be performed on Cereals data set My perception of understanding the given data set has been in the view of undertaking a client’s requirement of overcoming the stagnant point of sales of the products being manufactured by client.

To get a better understanding and work on a strategic approach for solution of the client, I have adapted the viewpoint of looking at ratings of the products and for further deep understanding of the problem, I have taken the stance of a consumer and reasoned out the various factors of choice of the products and they purchase , and my primary objective of this case study was to look up the factors which were dampening the sale of products and corelate them to ratings of products and draft out an outcome report to client regarding the various accepts of a product manufacturing , marketing and sale point determination.

vi

# Table of Contents:

LIST OF FIGURES IX

CHAPTER 1:MACHINE LEARNING

* 1. INTRODUCTION 1
  2. IMPORTANCE OF MACHINE LEARNING 1
  3. USES OF MACHINE LEARNING 2
  4. TYPES OF LEARNING ALGORITHMS 3
     1. Supervised Learning 3
     2. Unsupervised Learning 3
     3. Semi Supervised Learning 4
  5. [RELATION BETWEEN DATA MINING,MACHINE LEARNING AND DEEP LEARNING 5](#_TOC_250035)

[CHAPTER 2:PYTHON 6](#_TOC_250034)

* 1. [INTRODUCTION TO PYTHON 6](#_TOC_250033)
  2. [HISTORY OF PYTHON 6](#_TOC_250032)
  3. [FEATURES OF PYTHON 6](#_TOC_250031)
  4. [HOW TO SETUP PYTHON 7](#_TOC_250030)
     1. [Installation(using python IDLE) 7](#_TOC_250029)
     2. [Installation(using Anaconda) 8](#_TOC_250028)
  5. [PYTHON VARIABLE TYPES 9](#_TOC_250027)
     1. [Python Numbers 10](#_TOC_250026)
     2. [Python Strings 10](#_TOC_250025)
     3. [Python Lists 11](#_TOC_250024)
     4. [Python Tuples 11](#_TOC_250023)
     5. [Python Dictionary 12](#_TOC_250022)
  6. [PYTHON FUNCTION 12](#_TOC_250021)
     1. [Defining a Function 12](#_TOC_250020)
     2. [Calling a Function 13](#_TOC_250019)
  7. [PYTHON USING OOPs CONCEPTS 13](#_TOC_250018)
     1. Class 13
     2. [init method in class 14](#_TOC_250017)

CHAPTER 3:CASE STUDY 15

* 1. [PROBLEM STATEMENT 15](#_TOC_250016)
  2. [DATA SET 15](#_TOC_250015)
  3. [OBJECTIVE OF THE CASE STUDY 16](#_TOC_250014)

[CHAPTER 4:MODEL BUILDING 17](#_TOC_250013)

* 1. PRE PROCESSING OF DATA 17
     1. Getting the Dataset 18
     2. Importing the Libraries 19
     3. Importing the Dataset 20
     4. [Checking missing values in Dataset 22](#_TOC_250012)

CHAPTER 5: DATA PREPROCESSING / EDA 22

* 1. [STATISTICAL ANALYSIS 22](#_TOC_250011)
  2. [GENERATING PLOTS 23](#_TOC_250010)

[CHAPTER 6: SPLITTING OF DATA 28](#_TOC_250009)

* 1. [SPLITTING OF DATA 28](#_TOC_250008)
  2. [DATA SCALING 29](#_TOC_250007)

CHAPTER 7: MODEL BUILDING & EVALUATION 31

* 1. BRIEF ABOUT ALGORITHMS USED 31
     1. [Logistic Regression Model 31](#_TOC_250006)
     2. [Decision Tree Model 32](#_TOC_250005)
     3. [KNN Model 33](#_TOC_250004)
     4. Random Forest Model 34
     5. Grid Search CV on Random Forest Model 35
  2. [EVALUATION OF MODELS 36](#_TOC_250003)
  3. [MAKE PREDICTIONS 37](#_TOC_250002)
  4. PRICE RANGE PREDICTION FOR TEST.csv FILE 38

[CONCLUSION 39](#_TOC_250001)

[REFERENCES 40](#_TOC_250000)

viii

LIST OF FIGURES:

Figure 1 : The Process Flow 11

Figure 2 : Unsupervised Learning 12

Figure 3 : Semi Supervised Learning 13

Figure 4 : Python Download 16

Figure 5 : Anaconda Download 17

Figure 6 : Jupyter Notebook 17

Figure 7 : Importing Libraries 22

Figure 8 : Reading Dataset 23

Figure 9 : Missing Values in Dataset 24

Figure 10 : Statistical Analysis 24

Figure 11 : Heatmap of Dataset 25

Figure 12 : Joint plot 26

Figure 13 : Pie chart 27

Figure 14 : Bar graph 28

Figure 15 : Box plot 29

Figure 16 : Histogram 30

Figure 17 : Splitting of Dataset 31

Figure 18 : Data Scaling 32

Figure 19 : Logistic regression 33

Figure 20 : Decision tree 34

Figure 21 : KNN model 36

Figure 22 : Random Forest model 37

Figure 23 : Grid Search CV 38

Figure 24 : Prediction of Data 39

Figure 25 : Test.csv 40

Figure 26 : Predicted Dataset 42

CHAPTER 1

MACHINE LEARNING

* 1. INTRODUCTION:

Machine Learning(ML) is the scientific study of algorithms and statistical models that computer systems use in order to perform a specific task effectively without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of Artificial Intelligence(AI).

* 1. IMPORTANCE OF MACHINE LEARNING:

Consider some of the instances where machine learning is applied: the self-driving Google car, cyber fraud detection, online recommendation engines—like friend suggestions on Facebook, Netflix showcasing the movies and shows you might like, and “more items to consider” and “get yourself a little something” on Amazon—are all examples of applied machine learning. All these examples echo the vital role machine learning has begun to take in today’s data-rich world.

Machines can aid in filtering useful pieces of information that help in major advancements, and we are already seeing how this technology is being implemented in a wide variety of industries.

With the constant evolution of the field, there has been a subsequent rise in the uses, demands, and importance of machine learning. Big data has become quite a buzzword in the last few years; that’s in part due to increased sophistication of machine learning, which helps analyze those big chunks of big data. Machine learning has also changed the way data extraction, and interpretation is done by involving automatic sets of generic methods that have replaced traditional statistical techniques.

The process flow depicted here represents how machine learning works

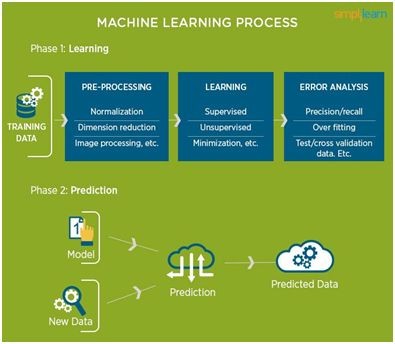


Figure 1 : The Process Flow

* 1. USES OF MACHINE LEARNING:

Earlier in this article, we mentioned some applications of machine learning. To understand the concept of machine learning better, let’s consider some more examples: web search results, real-time ads on web pages and mobile devices, email spam filtering, network intrusion detection, and pattern and image recognition. All these are by-products of applying machine learning to analyze huge volumes of data Traditionally, data analysis was always characterized by trial and error, an approach that becomes impossible when data sets are large and heterogeneous. Machine learning comes as the solution to all this chaos by proposing clever

alternatives to analyzing huge volumes of data. 2 By developing fast and efficient algorithms and data-driven models for real-time processing of data, machine learning can produce accurate results and analysis.

* 1. TYPES OF LEARNING ALGORITHMS:

The types of machine learning algorithms differ in their approach, the type of data they input and output, and the type of task or problem that they are intended to solve.

* + 1. Supervised Learning :

When an algorithm learns from example data and associated target responses that can consist of numeric values or string labels, such as classes or tags, in order to later predict the correct response when posed with new examples comes under the category of supervised learning.

Supervised machine learning algorithms uncover insights, patterns, and relationships from a labelled training dataset – that is, a dataset that already contains a known value for the target variable for each record. Because you provide the machine learning algorithm with the correct answers for a problem during training, it is able to “learn” how the rest of the features relate to the target, enabling you to uncover insights and make predictions about future outcomes based on historical data.

Examples of Supervised Machine Learning Techniques are Regression, in which the algorithm returns a numerical target for each example, such as how much revenue will be generated from a new marketing campaign.

Classification, in which the algorithm attempts to label each example by choosing between two or more different classes. Choosing between two classes is called binary classification, such as determining whether or not someone will default on a loan. Choosing between more than two classes is referred to as multiclass classification.

* + 1. Unsupervised Learning:

When an algorithm learns from plain examples without any associated response, leaving to the algorithm to determine the data patterns on its own. This type of algorithm tends to restructure the data into something else, such as new features that may represent a class or a new series of uncorrelated values. They are quite useful in providing humans with insights into the meaning of data and new useful inputs to supervised machine learning algorithms.

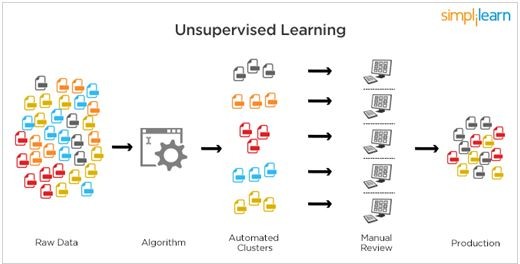


Figure 2 : Unsupervised Learning

Popular techniques where unsupervised learning is used also include self-organizing maps, nearest neighbor mapping, singular value decomposition, and k-means clustering. Basically, online recommendations, identification of data outliers, and segment text topics are all examples of unsupervised learning.

* + 1. Semi Supervised Learning:

As the name suggests, semi-supervised learning is a bit of both supervised and unsupervised learning and uses both labeled and unlabeled data for training. In a typical scenario, the algorithm would use a small amount of labeled data with a large amount of unlabeled data.

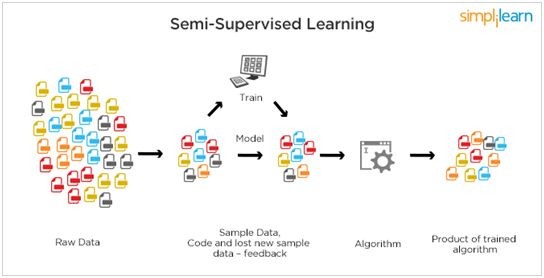


Figure 3 : Semi Supervised Learning

### RELATION BETWEEN DATA MINING,MACHINE LEARNING AND DEEP LEARNING:

Machine learning and data mining use the same algorithms and techniques as data mining, except the kinds of predictions vary. While data mining discovered previously unknown patterns and knowledge, machine learning reproduces known patterns and knowledge—and further automatically applies that information to data, decision-making, and actions.

Deep learning, on the other hand, uses advanced computing power and special 5 types of neural networks and applies them to large amounts of data to learn, understand, and identify complicated patterns. Automatic language translation and medical diagnoses are examples of deep learning.

CHAPTER 2 PYTHON

Basic programming language used for machine learning is : PYTHON

### INTRODUCTION TO PYTHON:

* Python is a high-level, interpreted, interactive and object-oriented scripting language.
* Python is a general purpose programming language that is often applied in scripting roles
* Python is Interpreted: Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is like PERL and PHP.
* Python is Interactive: You can sit at a Python prompt and interact with the interpreter directly to write your programs.
* Python is Object-Oriented: Python supports the Object-Oriented style or technique of programming that encapsulates code within objects.

### HISTORY OF PYTHON:

* + - Python was developed by GUIDO VAN ROSSUM in early 1990’s
* Its latest version is 3.7 , it is generally called as python3

### FEATURES OF PYTHON:

* + - Easy-to-learn: Python has few keywords, simple structure, and a clearly defined syntax, This allows the student to pick up the language quickly.
* Easy-to-read: Python code is more clearly defined and visible to the eyes.
* Easy-to-maintain: Python's source code is fairly easy-to-maintaining.
* A broad standard library: Python's bulk of the library is very portable and cross-platform compatible on UNIX, Windows, and Macintosh.
* Portable: Python can run on a wide variety of hardware platforms and has the same interface on all platforms.
* Extendable: You can add low-level modules to the Python interpreter. These modules enable programmers to add to or customize their tools to be more efficient.
* Databases: Python provides interfaces to all major commercial databases.
* GUI Programming: Python supports GUI applications that can be created and ported to many system calls, libraries and windows systems, such as Windows MFC, Macintosh, and the X Window system of Unix.

### HOW TO SETUP PYTHON:

* + - Python is available on a wide variety of platforms including Linux and Mac OS X. Let's understand how to set up our Python environment.
* The most up-to-date and current source code, binaries, documentation, news, etc., is available on the official website of Python.

### Installation(using python IDLE):

* + - * Installing python is generally easy, and nowadays many Linux and Mac OS distributions include a recent python.
* [Download python from www.python.org](http://www.python.org/)
* When the download is completed, double click the file and follow the instructions to install it.
* When python is installed, a program called IDLE is also installed along with it. It provides a graphical user interface to work with python.



Figure 4 : Python download

### Installation(using Anaconda):

* + - * Python programs are also executed using Anaconda.
* Anaconda is a free open source distribution of python for large scale data processing, predictive analytics and scientific computing.
* Conda is a package manager quickly installs and manages packages.
* In WINDOWS:
* In windows
  + Step 1: Open Anaconda.com/downloads in a web browser.
* Step 2: Download python 3.4 version for (32-bits graphic installer/64 -bit graphic installer)
* Step 3: select installation type( all users)
* Step 4: Select path(i.e. add anaconda to path & register anaconda as default python 3.4) next click install and next click finish
* Step 5: Open jupyter notebook ( it opens in default browser)



Figure 5 : Anaconda download

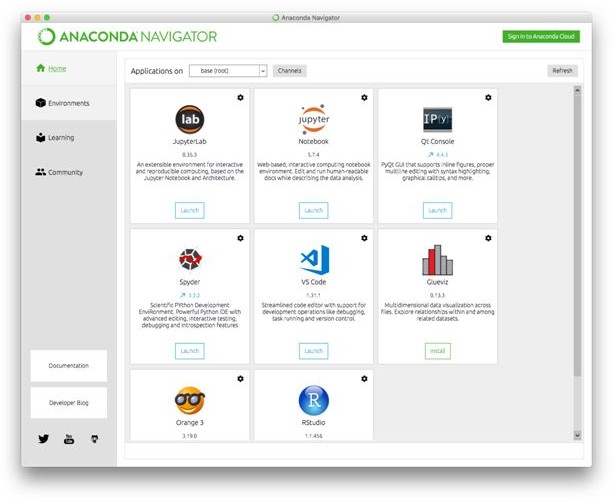


Figure 6 : Jupyter notebook

### PYTHON VARIABLE TYPES:

* + - Variables are nothing but reserved memory locations to store values. This means that when you create a variable you reserve some space in memory.
* Variables are nothing but reserved memory locations to store values.
* Based on the data type of a variable, the interpreter allocates memory and decides what can be stored in the reserved memory.
* Python variables do not need explicit declaration to reserve memory space. The declaration happens automatically when you assign a value to a variable.
* Python has various standard data types that are used to define the operations possible on them and the storage method for each of them.
* Python has five standard data types – o Numbers
* Strings
* Lists
* Tuples
* Dictionary

### Python Numbers:

* + - * Number data types store numeric values. Number objects are created when you assign a value to them.
* Python supports four different numerical types − int (signed integers) long (long integers, they can also be represented in octal and hexadecimal) float (floating point real values) complex (complex numbers).

### Python Strings:

* + - * Strings in Python are identified as a contiguous set of characters represented in the quotation marks.
* Python allows for either pairs of single or double quotes.
* Subsets of strings can be taken using the slice operator ([ ] and [:] ) with indexes starting at 0 in the beginning of the string and working their way from -1 at the end.
* The plus (+) sign is the string concatenation operator and the asterisk (\*) is the repetition operator.

### Python Lists:

* + - * Lists are the most versatile of Python's compound data types.
* A list contains items separated by commas and enclosed within square brackets ([]).
* To some extent, lists are similar to arrays in C. One difference between them is that all the items belonging to a list can be of different data type.
* The values stored in a list can be accessed using the slice operator ([ ] and [:]) with indexes starting at 0 in the beginning of the list and working their way to end -1.
* The plus (+) sign is the list concatenation operator, and the asterisk (\*) is the repetition operator.

### Python Tuples:

* A tuple is another sequence data type that is similar to the list.
* A tuple consists of a number of values separated by commas. Unlike lists, however, tuples are enclosed within parentheses.
* The main differences between lists and tuples are: Lists are enclosed in brackets ( [ ] ) and their elements and size can be changed, while tuples are enclosed in parentheses ( ( ) ) and cannot be updated.
* Tuples can be thought of as read-only lists.
* For example − Tuples are fixed size in nature whereas lists are dynamic. In other words, a tuple is immutable whereas a list is mutable. You can't add elements to a tuple. Tuples have no append or extend method. You can't remove elements from a tuple. Tuples have no remove or pop method.

### Python Dictionary:

* Python's dictionaries are a kind of hash table type. They work like associative arrays or hashes found in Perl and consist of key-value pairs. A dictionary key can be almost any Python type, but are usually numbers or strings. Values, on the other hand, can be any arbitrary Python object.
* Dictionaries are enclosed by curly braces ({ }) and values can be assigned and accessed using square braces ([]).
* You can use numbers to "index" into a list, meaning you can use numbers to find out what's in lists. You should know this about lists by now, but make sure you understand that you can only use numbers to get items out of a list.
* What a dict does is let you use anything, not just numbers. Yes, a dict associates one thing to another, no matter what it is.

### PYTHON FUNCTION:

### Defining a Function:

You can define functions to provide the required functionality. Here are simple rules to define a function in Python. Function blocks begin with the keyword def followed by the function name and parentheses (i.e.()).

Any input parameters or arguments should be placed within these parentheses. You can also define parameters inside these parentheses

The code block within every function starts with a colon (:) and is indented. The statement returns [expression] exits a function, optionally passing back an expression to the caller. A return statement with no arguments is the same as return None.

### Calling a Function:

Defining a function only gives it a name, specifies the parameters that are to be included in the function and structures the blocks of code. Once the basic structure of a function is finalized, you can execute it by calling it from another function or directly from the Python prompt.

### PYTHON USING OOP’s CONCEPTS:

* + 1. Class:
       - Class: A user-defined prototype for an object that defines a set of attributes that characterize any object of the class. The attributes are data members (class variables and instance variables) and methods, accessed via dot notation.
    - Class variable: A variable that is shared by all instances of a class. Class variables are defined within a class but outside any of the class's methods. Class variables are not used as frequently as instance variables are.
    - Data member: A class variable or instance variable that holds data associated with a class and its objects.
    - Instance variable: A variable that is defined inside a method and belongs only to the current instance of a class.
    - Defining a Class:

o We define a class in a very similar way how we define a function. o Just like a function ,we use parentheses and a colon after the class name(i.e. ():) when we define a class. Similarly, the body of our class is 14 indented like a functions body is indented like a functions body is.

### init method in Class:

* + - The init method — also called a constructor — is a special method that runs when an instance is created so we can perform any tasks to set up the instance.
* The init method has a special name that starts and ends with two underscores: init ().

## CHAPTER 3

CASE STUDY

### PROBLEM STATEMENT:

To predict the price range of the mobile phone with respect to its features by using Machine Learning classification algorithms

### DATA SET:

The given data set consists of the following parameters:

* 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 megapixels
* 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 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: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

### OBJECTIVE OF THE CASE STUDY:

This kind of prediction will help companies estimate the price of mobiles to give tough competition to other mobile manufacturers. Also it will be useful for Consumers to verify that they are paying the best price for a mobile.

## CHAPTER 4 MODEL BUILDING

### PREPROCESSING OF THE DATA:

Preprocessing of the data actually involves the following steps:

* + 1. GETTING THE DATASET: We can get the data set from the database or we can get the data from the client.
    2. IMPORTING THE LIBRARIES: We have to import the libraries as per the requirement of the algorithm.

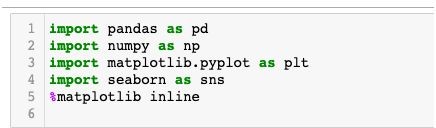
Importing Libraries:

Figure 7 : Importing Libraries

* + 1. IMPORTING THE DATA-SET:

Pandas in python provide an interesting method mobile.csv(). The read\_csv function reads the entire dataset from a comma separated values file and we can assign it to a DataFrame to which all the operations can be performed. It helps us to access each and every row as well as columns and each and every value can be accessed using the dataframe. Any missing value or NaN value has to be cleaned.

Reading the dataset:

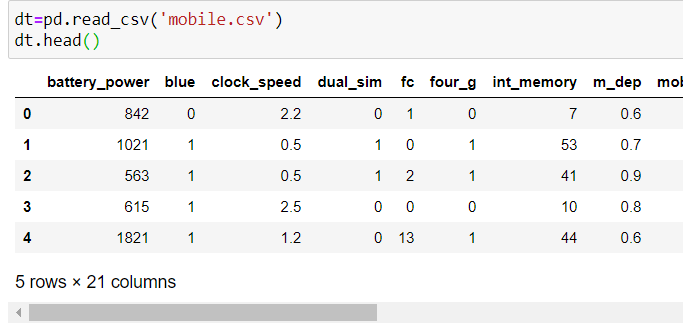


Figure 8 : Reading the dataset

### CHECKING MISSING VALUES IN DATASET:

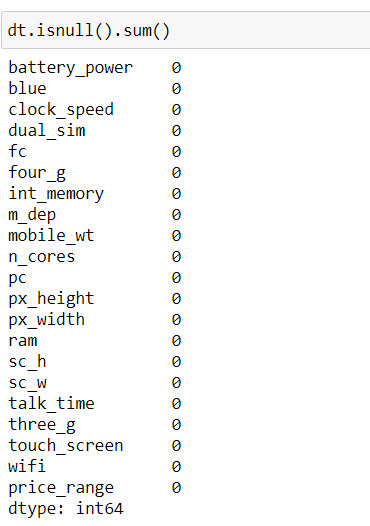
A 

Figure 9 : Missing values in dataset

## CHAPTER 5

DATA PREPROCESSING/FEATURE ENGINEERING AND EDA

### Statistical Analysis

Pandas DataFrame. describe() The describe() method is used for calculating some statistical data like percentile, mean and std of the numerical values of the Series or DataFrame. It analyzes both numeric and object series and also the DataFrame column sets of mixed data types.

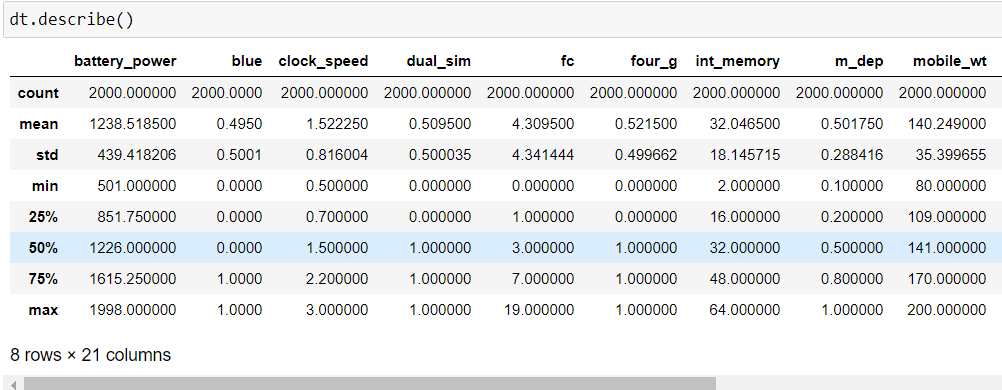


FIgure 10 : Statistical Analysis in dataset

### Generating Plots

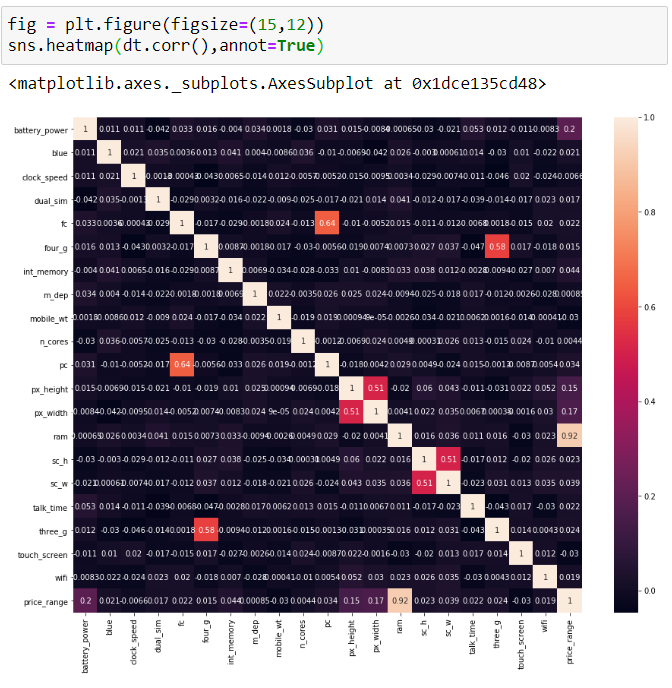
Visualize the data between Target and the Features

Data Correlation: Is a way to understand the relationship between multiple variables and attributes in your dataset. Using Correlation, you can get some insights such as: One or multiple attributes depend on another attribute or a cause for another attribute.

* A heat map (or heatmap) is a data visualization technique that shows the magnitude of a

phenomenon as color in two dimensions. The variation in color may be by hue or intensity, giving obvious visual cues to the reader about how the phenomenon is clustered or varies over space.

Figure 11: Heatmap of the dataset



* By the above heatmap we can infer that Ram feature is highly correlated .
* seaborn. jointplot. Seaborn's jointplot displays a relationship between 2 variables (bivariate) as well as 1D profiles (univariate) in the margins. This plot is a convenience class that wraps JointGrid.

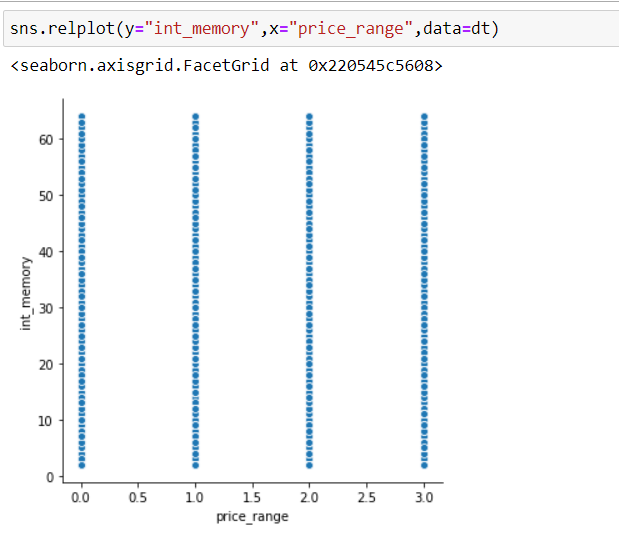


Figure :12 Jointplot of dataset

* The above joint plot shows the relation between the features memory and price range.
  + In the relational plot we saw how to use different visual representation to show the relationship between muiltiple variables in a dataset. The above is an example.

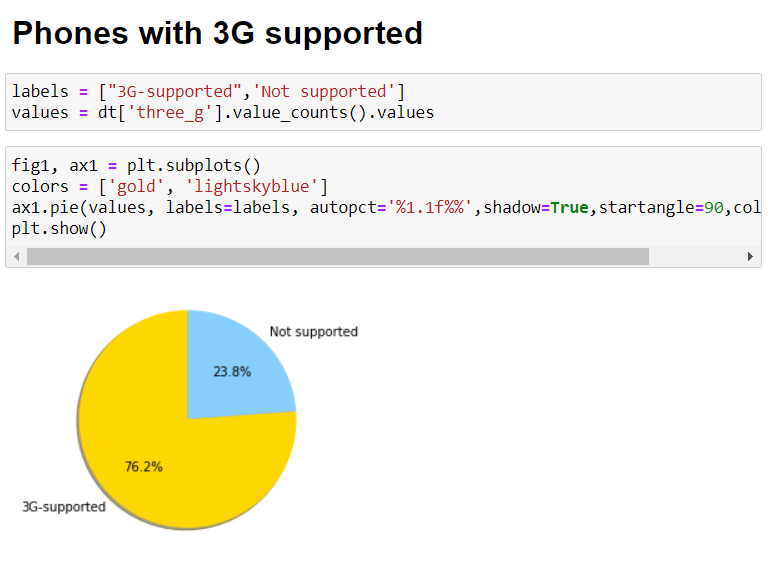


Figure :13 Pie charts of the dataset

* A bar graph shows comparisons among discrete categories. One axis of the chart shows the specific categories being compared, and the other axis represents a measured value.



Figure :15 Boxplot

* A histogram is a graphical display of data using bars of different heights. In a histogram, each bar groups numbers into ranges. Taller bars show that more data falls in that range. A histogram displays the shape and spread of continuous sample data.

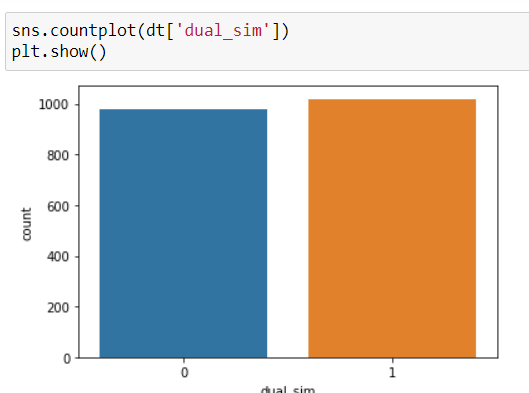


Figure :14 Bar Graph of dataset

* In descriptive statistics, a box plot or box plot is a method for graphically depicting groups of numerical data through their quartiles. Box plots may also have lines extending from the boxes (whiskers) indicating variability outside the upper and lower quartiles, hence the terms box-and-whisker plot and box-and-whisker diagram. Outliers may be plotted as individual points.

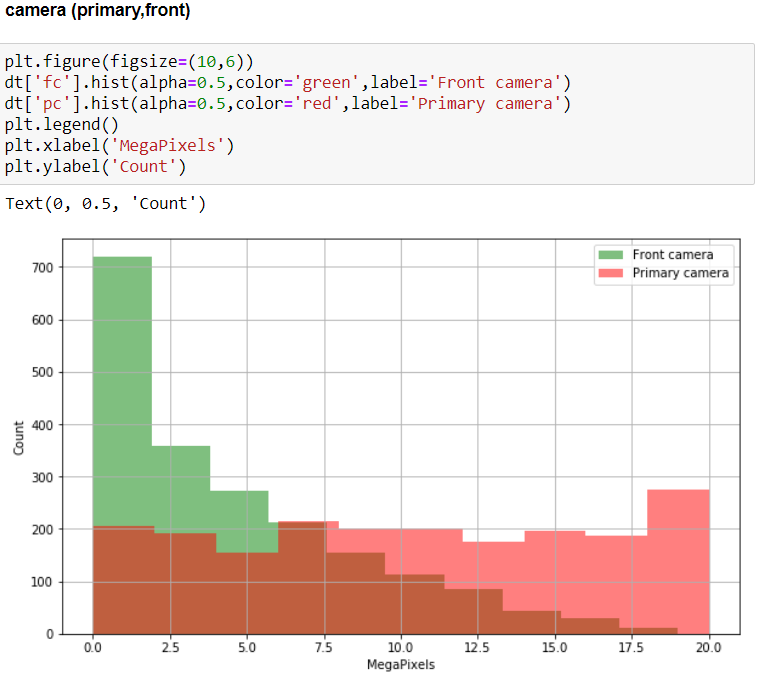


Figure : histogrms

## CHAPTER 6 SPLITTING OF DATA

### Splitting Of Data:

* Splitting the data : after the preprocessing is done then the data is split into train and test sets
* In Machine Learning in order to access the performance of the classifier. You train the classifier using 'training set' and then test the performance of your classifier on unseen 'test set'. An important point to note is that during training the classifier only uses the training set . The test set must not be used during training the classifier. The test set will only be available during testing the classifier.
* training set - a subset to train a model.(Model learns patterns between Input and Output)
* test set - a subset to test the trained model.(To test whether the model has correctly learnt )
* The amount or percentage of Splitting can be taken as specified (i.e. train data = 75% , test data =25% or train data = 80% , test data= 20%)
* First we need to identify the input and output variables and we need to separate the input set and output set
* In scikit learn library we have a package called model\_selection in which train\_test\_split method is available .we need to import this method
* This method splits the input and output data to train and test based on the percentage specified by the user and assigns them to four different variables(we need to mention the variables)

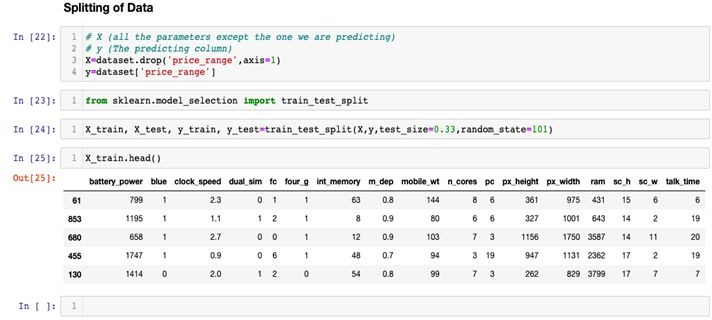


Figure :18 Splitting of Dataset

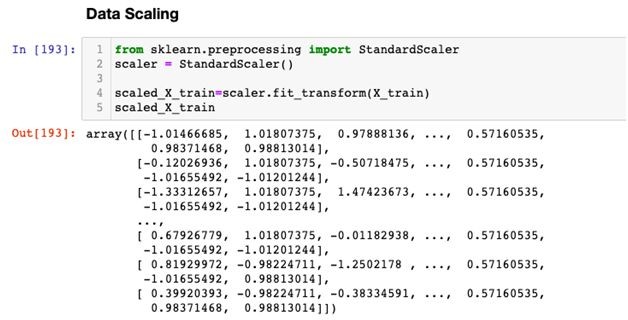
### Data Scaling

Feature scaling (also known as data normalization) is the method used to standardize the range of features of data. Since, the range of values of data may vary widely, it becomes a necessary step in data preprocessing while using machine learning algorithms.

The idea behind StandardScaler is that it will transform your data such that its distribution will have a mean value 0 and standard deviation of 1.

In case of multivariate data, this is done feature-wise (in other words independently for each column of the data).

Given the distribution of the data, each value in the dataset will have the mean value subtracted, and then divided by the standard deviation of the whole dataset (or feature in the multivariate case).



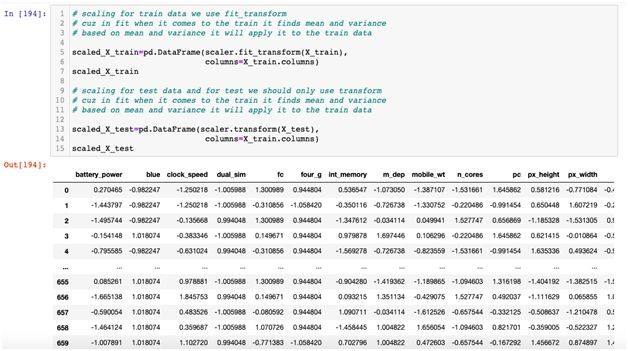


Figure: 19 Data Scaling

## CHAPTER 7

MODEL BUILDING AND EVALUATION

### Brief about the algorithms used

### Logistic regression Model

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of the target or dependent variable is dichotomous, which means there would be only two possible classes.

In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.

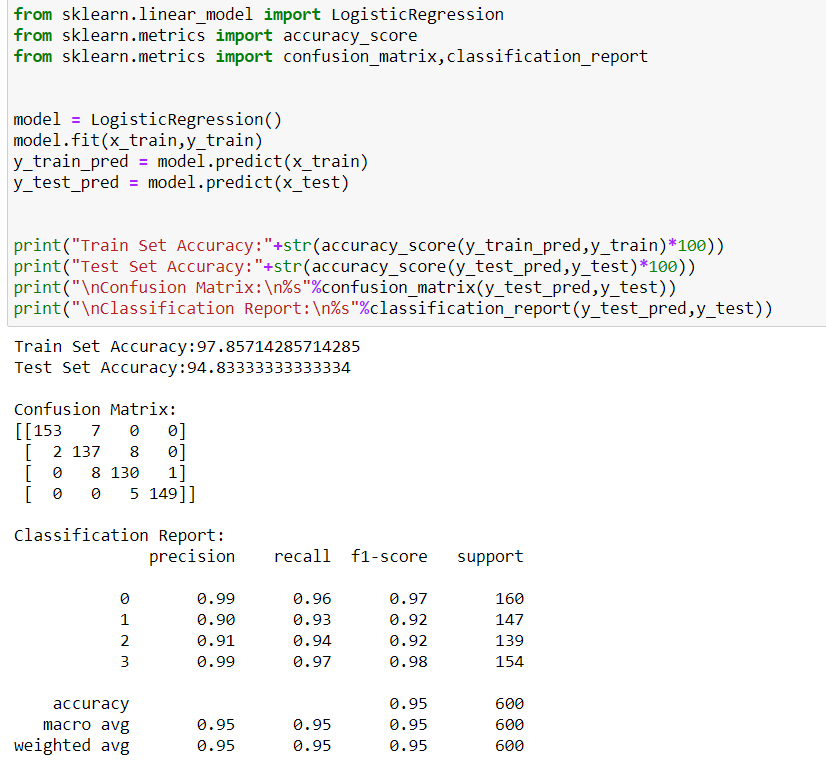


Figure: 20 Logistic Regression

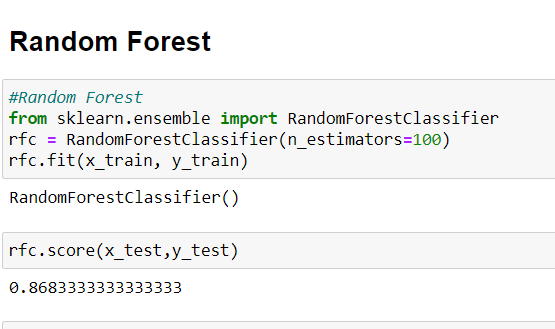
### Decision Tree Model

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

A decision tree is a flowchart-like structure in which each internal node represents a “test” on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

Figure : 21 Decision tree

### Random forest Model



### Random Forest Model

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

The first algorithm for random decision forests was created by Tin Kam Ho using the random subspace method, which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg.

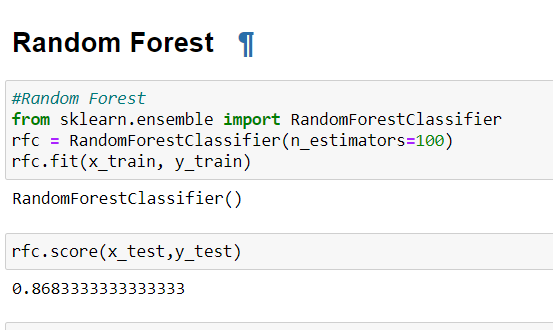


Figure: 23 Random Forest Model

### Grid Search CV

Grid search is an approach to parameter tuning that will methodically build and evaluate a model for each combination of algorithm parameters specified in a grid.

7.1.5 Grid Search CV On Random Forest Algorithm

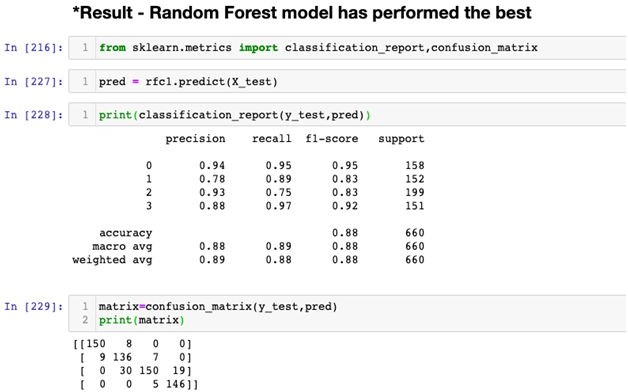
### Evaluation of Models

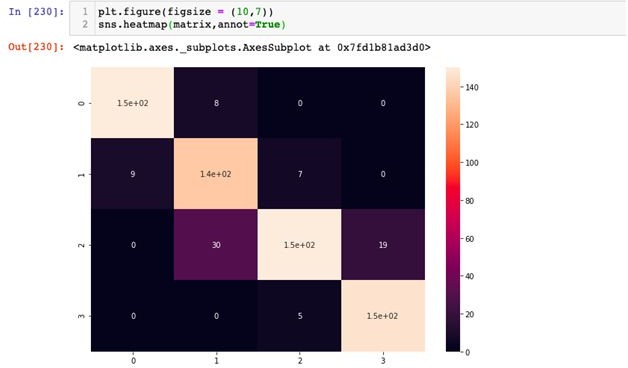
Here we have used 4 models i.e

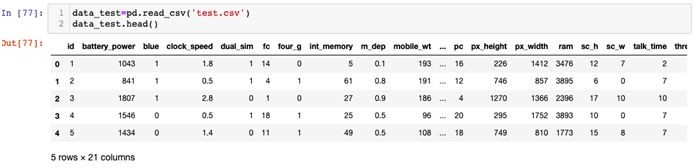
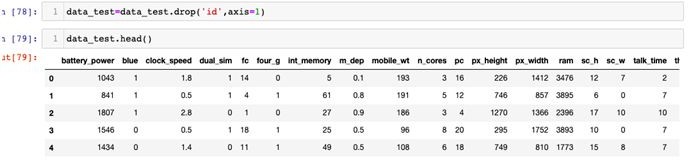
* Logistic regression
* Decision Tree
* Random Forest Model

So by comparing all the accuracies we can infer that the Random Forest Classifier Algorithm model has a good accuracy ,so the Random Forest Classifier model is selected among the models for predicting the price range.

### Make Predictions



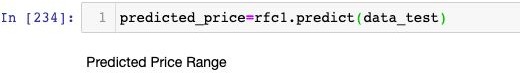


Figure: 25 Prediction of data

### Price Range Prediction for Test.csv using Random Forest Model

* Importing the Test.csv data

### Prediction using Model



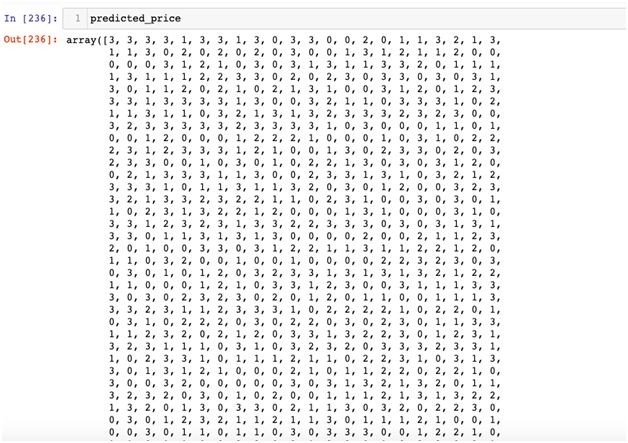


Figure: 26 predicted price

### Adding the Predicted Price Range to the Dataset

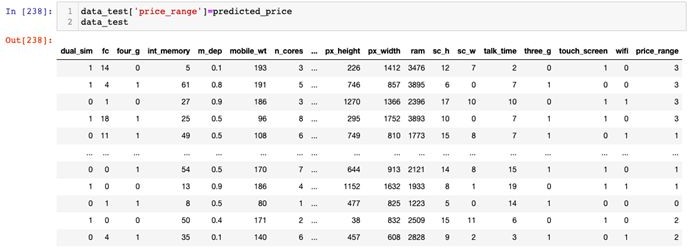


Figure: 27 Predicted Dataset.

## CONCLUSION:

For designing the model for predicting the price range , I have applied Logistic Regression, Decision Tree, Random Forest, Grid Search CV. Among these models Random Forest Model has performed well and the price prediction of test data has done with Random Forest Model.

The classification of each data point of test.csv is done into either of 4 classes (0,1,2,3)