**PROFESSIONAL TRAINING REPORT**

**at**

**Sathyabama Institute of Science and Technology (Deemed to be University)**

Submitted in partial fulfillment of the requirements for the award of

Bachelor of Engineering Degree in Computer Science an Engineering

By

**PASHAM VYSHNAVI**

**(Reg. No :40111456)**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING SCHOOL OF COMPUTING SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI – 600119, TAMILNADU**

**APRIL 2023**

**SATHYABAMA**

**INSTITUTE OF SCIENCE AND TECHNOLOGY**

**(DEEMEDTOBEUNIVERSITY)**

**Accredited with Grade “A” by NAAC**

**(Established under Section 3 of UGC Act, 1956)**

**JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI– 600119**

[**http://www.sathyabama.ac.in**](http://www.sathyabama.ac.in/)

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**BONAFIDE CERTIFICATE**

This is to certify that this Project Report is the bonafide work of **PASHAM VYSHNAVI (40111456)** who carried out the project entitled “**BANK MARKETING DATASET**” under my supervision from Jan 2023 to March 2023.

**Internal Guide**

**Dr. S.Bangaru Kamatchi M.E.,Ph.D.,**

**Head of the Department**

**Dr.L.Lakshmanan,M.E.,Ph.D.**

**Submitted for Viva voce Examination held on**

**Internal Examiner External Examiner**

**DECLARATION**

I**, PASHAM VYSHNAVI** hereby declare that the Project Report entitled **“BANK MARKETING DATASET”** by me under the guidance of is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering Degree in Computer Science and Engineering.

**DATE:**

**PLACE: SIGNATURE OF THE CANDIDATE**

**ACKNOWLEDGEMENT**

I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. Sasikala M.E., Ph.D.**, **Dean**, School of Computing, **Dr.L.Lakshmanan M.E., Ph.D.,** Heads of the Department of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr.S.Bangaru Kamatchi M.E.,Ph.D.,** for his valuable guidance, suggestions and constant encouragement paved the way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

**TRAINING CERTIFICATE**

**ABSTRACT**

In today’s world, where huge amount of data is generated in every field of day-to-day activities, banking sector is one of them. As an outcome of work, various machine learning concept are studied with respect to Bank marketing data classification. Banking is a provision of the services by bank to an individual customer. The dataset is originally collected from UCI Machine learning repository and Kaggle website. The data is related to bank marketing campaigns of banking institution based on phone call. In this work, Python is used as a coding language and Machine learning concept is used as statistical learning for data analysis. The main reason of using machine learning is to build a predictive model to produce the better prediction. The outcome of the result is analyzed with supervised Naïve Bayes algorithm for classification purpose. The main objective of building the model is to describe whether the customer has opted for term deposit. The bank should target the potential customer with considerable amount of time responding to the phone calls. The work implemented resulted in measuring accuracy, precision, recall and F1 score, towards term deposit prediction.This bank marketing dataset project is implemented by using logistic regression and decision tree algorithm Logistic regression model presents the relationship between quantitative feature. The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (variable y). Data from a marketing campaign run by Bank code Portugal is examined. The campaign’s aim was to increase customers’ subscription rates to fixed-term deposit products, such as CDs.

**TABLE OF CONTENTS**

**CHAPT TITLE PAGE NO**

**ER No.**

ABSTRACT

LIST OF FIGURES

LIST OF ABBREVIATIONS

1. **INTRODUCTION**
   1. BANK MARKETING DATASET 1
   2. ESTIMATING DATASET INFORMATION 2
2. **AIM AND SCOPE OF THE PRESENT**

**INVESTIGATIONAL**

* 1. AIM 5
  2. SYSTEM REQUIREMENTS 5
  3. SYSTEM ARCHITECTURE 6

1. **EXPERIMENTAL OR MATERIALS AND**

**METHODS;**

**ALGORITHM USED**

* 1. ALGORITHM 7
  2. TYPES OF CLASSIFICATION

ALGORITHM USED 10

1. **RESULTS AND DISCUSSION,**

**PERFORMANCE ANALYSIS**

* 1. SOFTWARE REQUIREMENTS 16
  2. RESULT 17

1. **SUMMARY AND CONCLUSION**
   1. CONCLUSION 18

**REFERENCES**

**APPENDIX**

**SCREENSSHOTS** 19

**LIST OF FIGURES**

**S.NO FIGURE NO FIGURE NAME PAGE NO**

1. **Fig :1.2.1 data set information 02**
2. **Fig:2.3.1 architecture 06**
3. **Fig:3.1.1 LR Graph 07**
4. **Fig:3.2.2 Python Modules 10**
5. **Fig:3.2.3 Graph 11**
6. **Fig:3.2.4 Seaborn Graph 13**
7. **Fig:3.3.1 Decision Tree Graph 15**

**8 Fig:4.4.1 Anaconda console 16**

**9 Fig:4.1.2 Jupyter Console 17**

**LIST OF ABBREVATIONS**

**ABBREVATION EXPANSION**

ML MACHINE LEARNING

AI ARTIFICIAL INTELLIGENCE

Np NUMPY

Pd PANDAS

sklearn SCIKIT LEARN

NUMPY NUMERICAL PYTHON

**CHAPTER-1**

**INTRODUCTION**

## 1.1 BANK MARKETING DATASET

In machine learning and statistics, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, based on a training set of data containing observations (or instances) whose category membership is known. we will use the Bank Marketing Dataset from Kaggle to build a model to predict whether someone is going to make a deposit or not depending on some attributes. We will try to build both logistic regression and decision tree algorithm. After building these models we will evaluate them and compare with other models this is best for our case. We will then try to optimize our model by tuning the hyper parameters of the model by using Grid Search, Lastly we will save . the prediction result from our dataset and then save our model for reusability. To start we will load some basic libraries such as NumPy and Pandas then make some configuration to some of those libraries

The Bank marketing dataset has 41188 examples with 20 inputs and 1 output variable. This data is from Portuguese Banking institution. It has numerical as well as categorical attributes and response attribute y denotes client subscribed to term deposit or not .The goal is to build models that can predict if client will subscribe to term deposit or not. Since response variable is binary, different classification models will be used incrementally till it gives model with best accuracy The process by which companies create value for customers and build strong customer relationships in order to capture value from customers in return. Kotler and Armstrong (2010).Marketing campaigns are characterized by focusing on the customer needs and their overall satisfaction. Nevertheless, there are different variables that determine whether a marketing campaign will be successful or not. There are certain variables that we need to take into consideration when making a marketing campaign.

## 1.2 ESTIMATING DATASET INFORMATION

The given dataset contains 45212 rows and 17 coloumns.

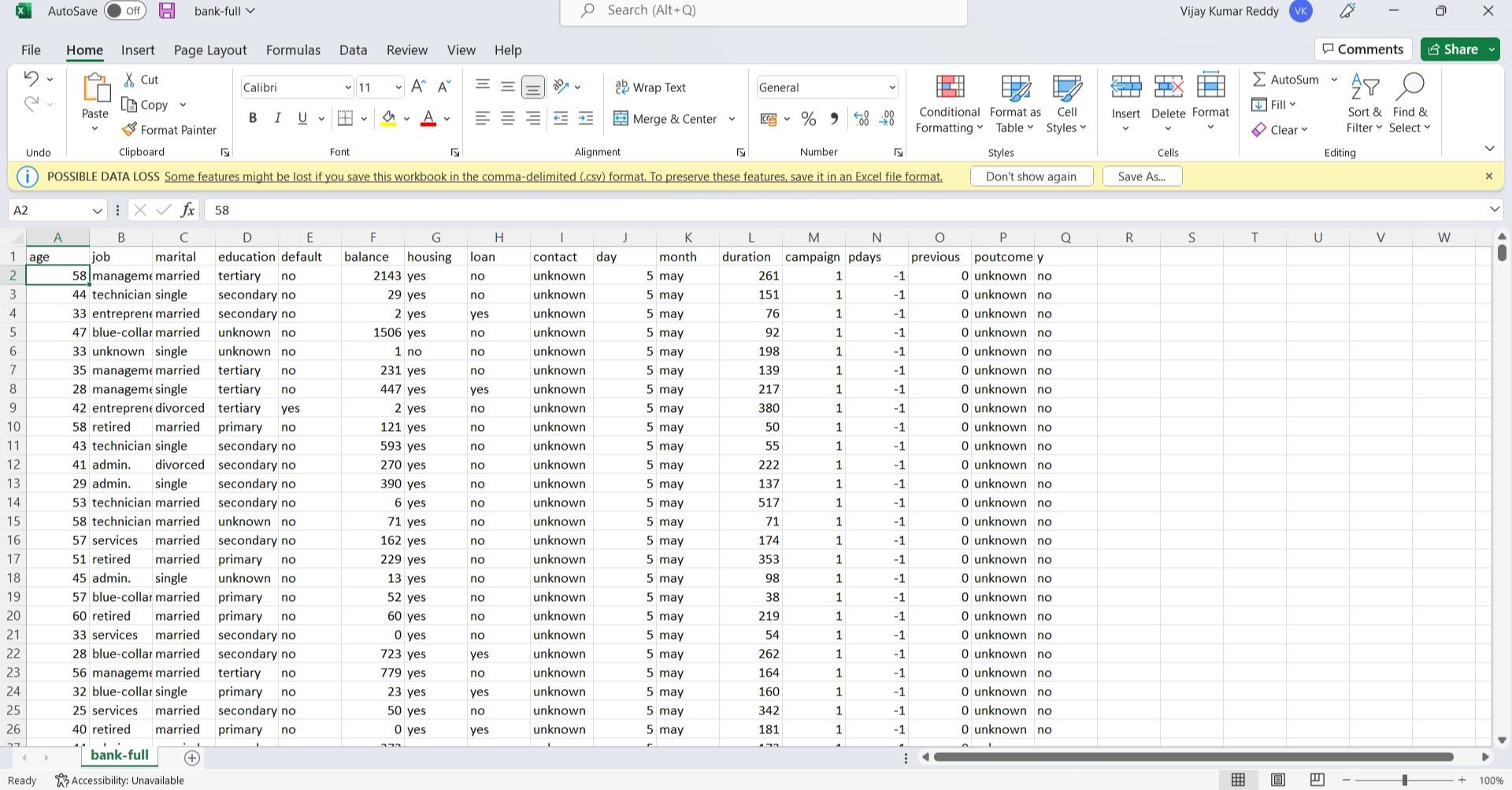


Fig :1.2.1 data set information

**1.2.1 Attribute Information:**

Input variables:

1. - age (numeric)
2. - job: type of job (categorical: 'admin.', 'blue

collar','entrepreneur','housemaid','management','retired','selfemployed','services','student','technician','unemployed','unknown')

1. – Marital : marital status (categorical: 'divorced','married','single','unknown';)

4- education(categorical:'basic.4y','basic.6y','basic.9y','high.school','illiterate',

'professional.course','university.degree','unknown')

1. - default: has credit in default? (categorical: 'no','yes','unknown')
2. - balance : current balance in account
3. - housing: has housing loan? (categorical: 'no', 'yes', 'unknown')
4. - loan: has personal loan? (categorical: 'no','yes','unknown')
5. - contact: contact communication type (categorical: 'cellular','telephone') 10 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')

11-day\_of\_week: last contact day of the week (categorical:

'mon','tue','wed','thu','fri')

1. - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no').

Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known.

1. - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
2. - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
3. - previous: number of contacts performed before this campaign and for this client (numeric)
4. - poutcome: outcome of the previous marketing campaign (categorical:

'failure','non existent','success')

Output variable (desired target):

1. - y - has the client subscribed a term deposit? (binary: 'yes','no')4

**1.2.2 Data Set Information:**

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

There are four datasets:

1. bank-additional-full.csv with all examples (45212) and 20 inputs, ordered by date very close to the data analyzed .
2. bank-additional.csv with 10% of the examples (45212), randomly selected from 1), and 20 inputs.
3. bank-full.csv with all examples and 17 inputs, ordered by date (older version of this dataset with less inputs).
4. bank.csv with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs).

The smallest datasets are provided to test more computationally demanding machine learning algorithms (e.g., SVM).

**CHAPTER 2**

## 2.1 AIM

The main of the project is project to predict the value in the given dataset(bank marketing) using machine learning. To predict the y values in the given bank marketing dataset using features in the given dataset To find the solution we used logistic regression and decision tree algorithm

## 2.2 SYSTEM REQUIREMENTS

**2.2.1 SOFTWARE REQUIREMENT**

|  |  |  |
| --- | --- | --- |
| Operating System | - | Windows OS |
| Coding Language | - | Python |
| Tool | - | Jupyter Notebook |
| Libraries | - | numpy,pandas,matplotlib, |

seaborn

#### **2.2.2 HARDWARE REQUIREMENT**

The following hardware requirement are

CPU : 2 x 64\_bit 2.8 GHz 8.00 GT/s CPUs

RAM :32 GB (or 16 GB of 1600 MHz DDR3 RAM)

## 2.3 System Architecture:

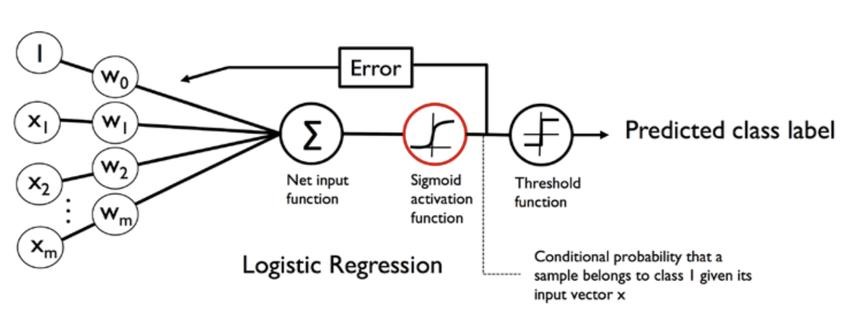


Fig:2.3.1 ARCHITECTURE

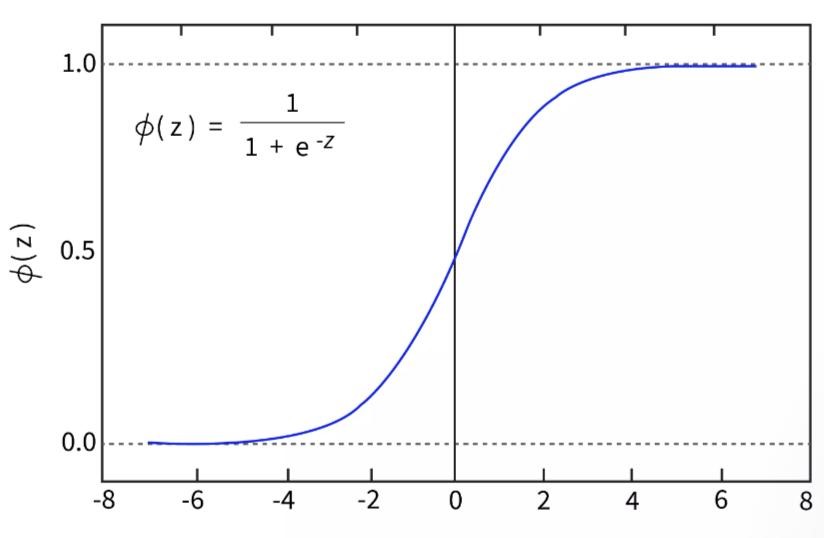
**CHAPTER – 3 EXPERIMENTAL OR MATERIALS AND METHODS**

## 3.1 ALGORITHMS

**BOTH LOGISTIC REGRESSION AND DECISION TREE**

**3.1.1.Logistic Regression :**

|  |
| --- |
| Logistic Regression Machine Learning is basically a classification algorithm that comes under the Supervised category (a type of machine learning in which |
| machines are trained using "labelled" data, and on the basis of that trained data, |
| the output is predicted) of Machine Learning algorithms. The main role of Logistic Regression in Machine Learning is predicting the output of a categorical dependent variable from a set of independent variables. In simple words, categorical dependent variable means a variable that is dichotomous or binary in nature having its data coded in the form of either 1 (stands for success/yes) or 0 (stands for failure/no). Although Logistic Regression is one the simplest machine learning algorithms, it has got diverse applications in classification problems ranging from spam detection, diabetes prediction to even cancer |



detection.

FIG 3.1.1 LR graph

**3.1.2 Logistic regression analysis of bank marketing dataset:**

1.Binary logistic regression requires the dependent variable to be binary.

1. For a binary regression, the factor level 1 of the dependent variable should represent the desired outcome.
2. Only the meaningful variables should be included.
3. The independent variables should be independent of each other. That is, the model should have little or no multicollinearity.

5 .The independent variables are linearly related to the log odds.

6. Logistic regression requires quite large sample sizes.

#### **3.1.3 Data**

The dataset comes from the [UCI](http://archive.ics.uci.edu/ml/index.php) [Machine](http://archive.ics.uci.edu/ml/index.php) [Learning](http://archive.ics.uci.edu/ml/index.php) [repository](http://archive.ics.uci.edu/ml/index.php), and it is related to direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict whether the client will subscribe (1/0) to a term deposit (variable y).

|  |
| --- |
| import pandas as pd import numpy as np  from sklearn import preprocessing import matplotlib.pyplot as plt plt.rc("font", size=14)  from sklearn.linear\_model import LogisticRegression from sklearn.model\_selection import train\_test\_split import seaborn as sns sns.set(style="white")  sns.set(style="whitegrid", color\_codes=True) |

The data set provides the bank customers information It includes 45212 records and 17 fields.

**3.1.4 Methodology :**

1 .In machine learning we used classification method, logistic regression

### algorithm.

1. .Technology: Jupyter is a free, open-source, interactive web tool known as a computational notebook, which researchers can use to combine software code, computational output, explanatory text and multimedia resources in a single document.
2. .Software tools:

-Machine learning

-python3

1. . By using above technology ,software tools we are going to achieve our objective.

**3.1.5 Logistic regression algorithm :**

1.Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

2.Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

3.Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.

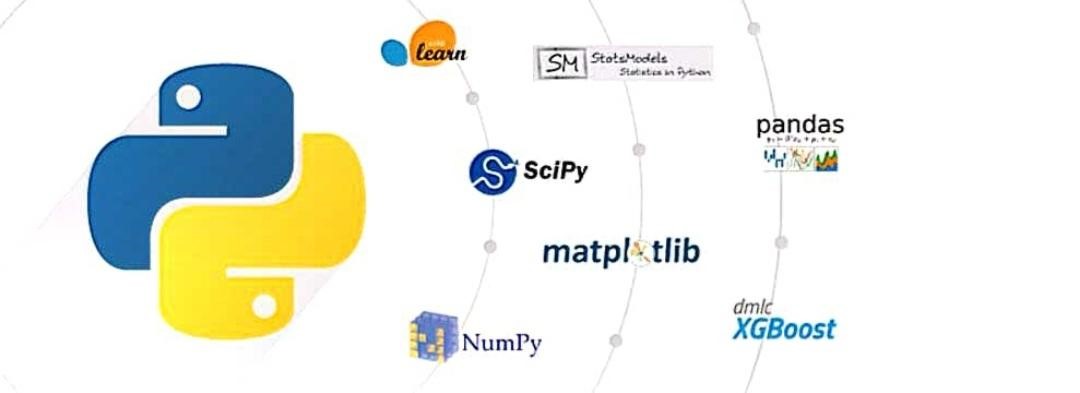
**3.2 Types of classification algorithm used:**

**3.2.1 Decision tree Algorithm:**

• Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. • It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. 20 • In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node • It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

**3.2.2 IMPORTED LIBRARIES :**

Libraries are collections of prewritten code that users can use to optimize tasks. In project as python is used for implementation tool, it has the most libraries as compared to other programming languages. More than of 60% machine learning developers use and goes for python as it is easy to learn. As python has comparatively large collection of libraries let’s look at the libraries that came in handy for mammographic dataset.



**FIG 3.2.2 pyhton modules**

**3.2.3 LIBRARIES USED:**

**3.2.3.1 Numpy** stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of countless of routines for processing those arrays. Using this mathematical and logical operations on arrays can be performed. The 21 difference in using Numpy from pandas is, it works on numerical data whereas pandas on tabular data

**3.2.3.2 Pandas** is a widely-used data analysis and manipulation library for python. It provides a lot of functions and methods that expedite the data analysis and preprocessing steps. IT also provides fast, flexible and expressive data structures working with relational or labelled or both easy and intuitive. Considered as fundamental high-level building block in performing practical, realworld data analysis in python. Has powerful tools like Data Frame and Series for analysing

**3.2.3.3 Matplotplib.pyplot** is a state-based interface to matplotlib. It provides a MATLAB-like way of plotting. It makes changes to figures.

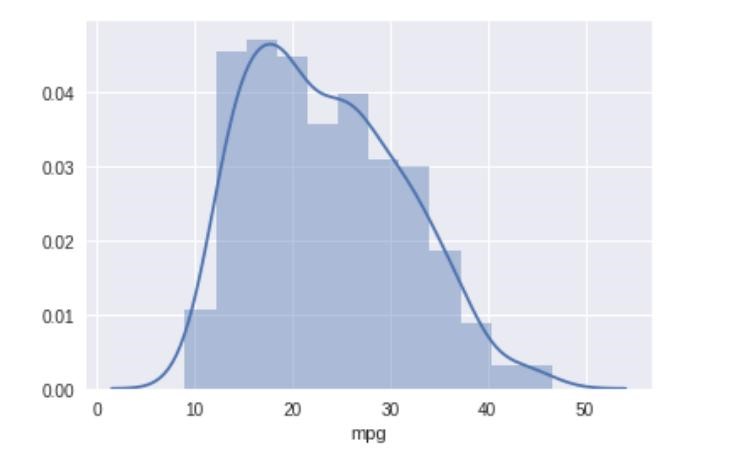


Fig:3.2.3 Graph

**3.2.3.4 sklearn** stands for Scikit-learn, a machine learning library. It is imported for various classification, regression and clustering algorithms including k-means, random forest, support vector machines, gradient boosting and DBSCAN. It is designed using libraries Numpy and Scipy. From the sklearn library and from the tree inside the library DecisionTreeClassifier. It is a class capable of performing multi-class classifier on a dataset. When compared with other classifiers, DecisionTreeClassifier takes input as two arrays: an array X, a parse or dense, of shape (n\_samples, 22 n\_features) holding training samples and an array Y of integer values, shape (n\_samples), holding class labels for training sample. From sklearn another one called model\_selection for training and testing the model imports train\_test\_split. It is a method setting a blueprint to analyse data and the using it to measure new data. Selecting a proper model allows to generate accurate results while making prediction. For proceeding, we need to train the model by using a specific dataset and test the model against another dataset. By default, sklearn train\_test\_split will make random partitions for two subsets. We can also specify a

random state for the operation. First, we need to split the dataset and then allocate the size for train and test. For this mammographic dataset we need train size as 80% (0.80) and test size as 20% (0.20) with the random state of 100.

**3.2.3.5. Operator** module exports a set of efficient functions corresponding to the intrinsic operators of Python. For example, operator. add (x, y) is equivalent to the expression x+y. Many function names are those used for special methods, without the double underscores. For backward compatibility, many of these have a variant with the double underscores kept. The variants without the double underscores are preferred for clarity. The functions fall into categories that perform object comparisons, logical operations, mathematical operations and sequence operations.

**3.2.3.6.Seaborn** is a library built on top of matplotlib. It used for data visualization and exploratory data analysis. They work easily with dataframes and panda’s library. The graphs created can also be customized easily. It provides default styles and colour palettes to make statistical plots more attractive.

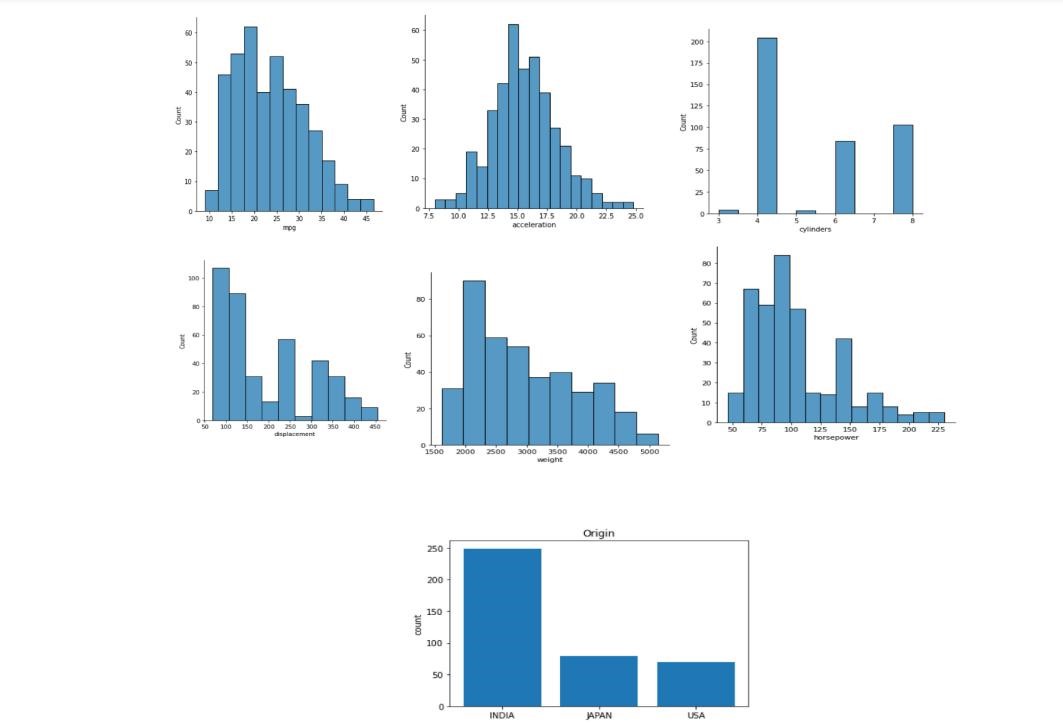


Fig:3.2.4 seaborn graphs

**3.2.3.7 Implementation of Decision Tree:**

1. Import the packages and classes you need.
2. Provide data to work with and eventually do appropriate transformations.
3. Create a regressor model and fit it with existing data.
4. Check the results of model fitting to know whether the model is satisfactory.
5. Apply the model for predictions.

#### **Fitting the values into Decision Tree Regressor**



**3.3.1.DECISION TREE:** Decision Tree is the most powerful and popular tool for classification and Fig 1.3.3- Decision Tree scatter plot. 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. 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 17 or values. here it is important to continuous values are predicted with the help of a decision tree regression

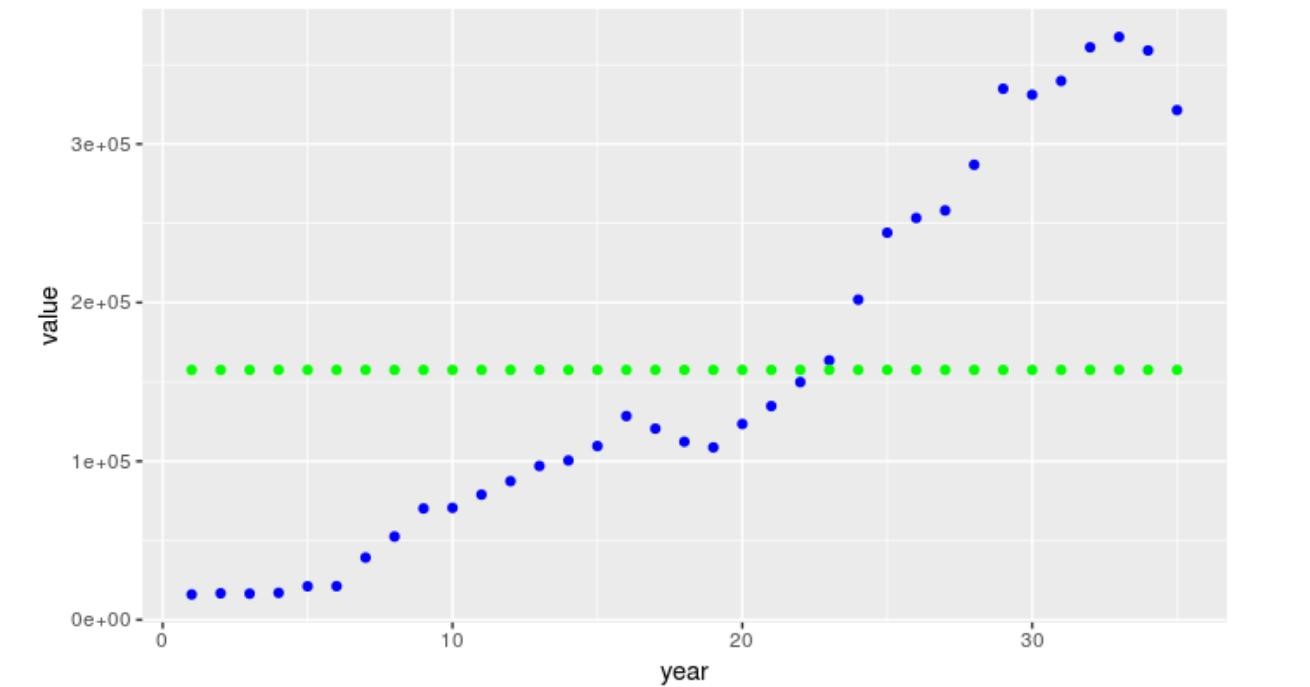


FIG:3.3.1 decision tree graph

Decision tree model 1

Also, the models take some understanding as to how to tune them. The above graph captures the data only at one point - at about 23 years in, the model is perfect, but all other points are completely ignored. In this situation, you would be better off averaging all the data points to get your prediction

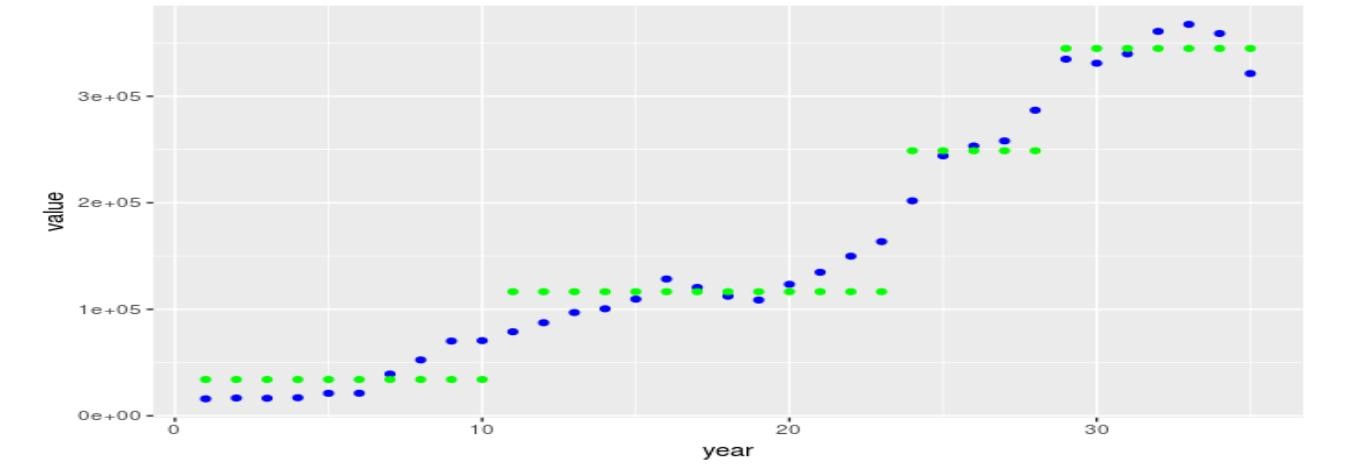


Fig:3.3.2

Decision Tree Model 2

**CHAPTER 4**

# 4.1 RESULTS AND DISCUSSION, PERFORMANCE ANALYSIS

## 4.1.1 SOFTWARE REQUIREMENTS:

ANACONDA NAVIGATOR is desktop GUI used to launch applications and also manage packages in one place. Outlook

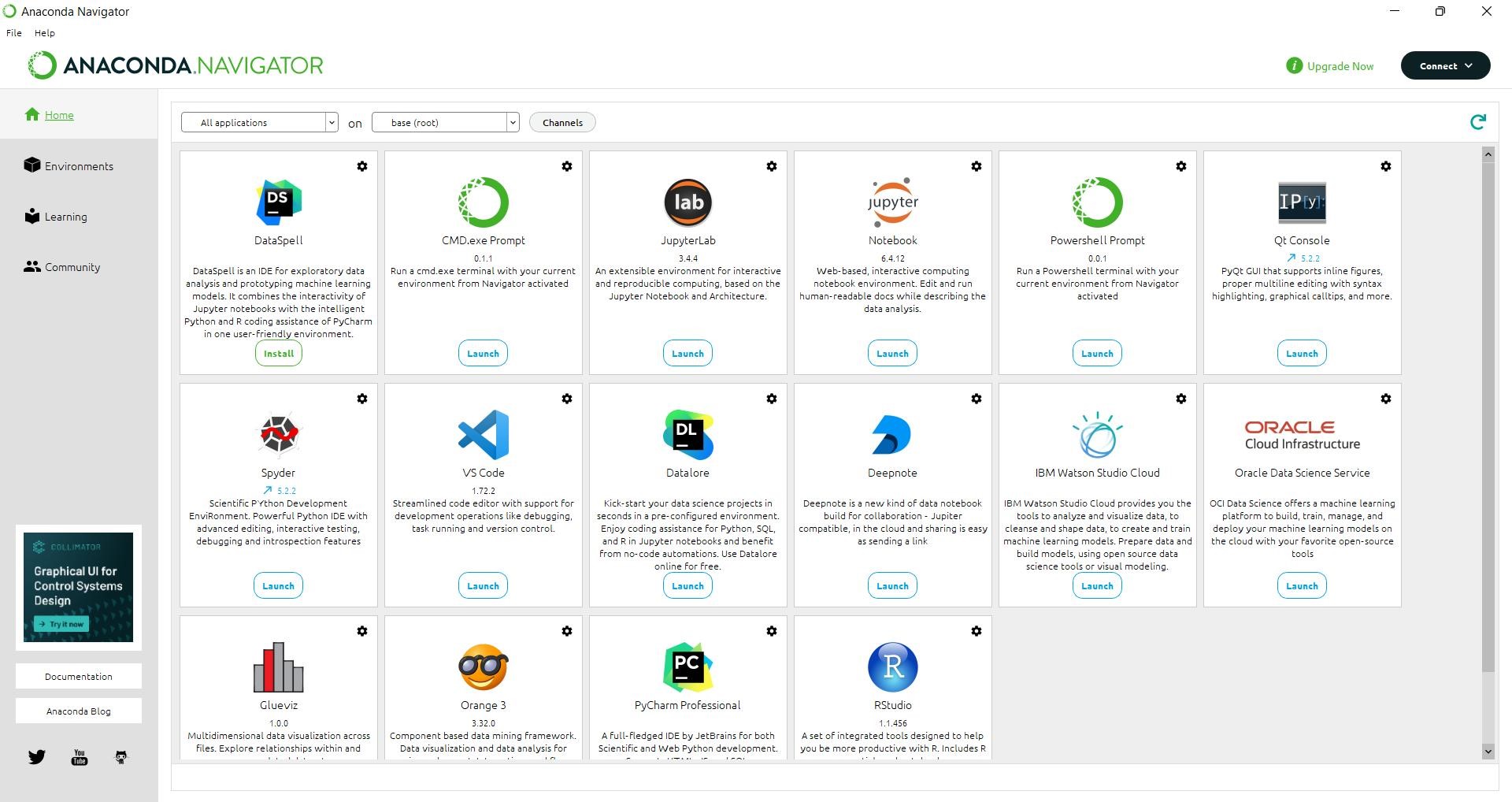


Fig:4.1.1 Anaconda console

**4.1.2 Coding Environment :**

Jupyter notebook from the anaconda navigator is launched along with all the preinstalled packages for python.



**FIG 4.1.2 jupyter console**

## 4.2 RESULT

From the feature importance plot it can be inferred attributes such as age, job, marital status, education, default, housing, current balance, contact information, duration ,campaign pdays, outcome are some of the most important features in predicting the outcome y value. The object of the project is get y value and better accuracy. The project accuracy is 88%.

**CHAPTER 5**

**5.1 SUMMARY AND CONCLUSION**

This sums up for the classification task of bank marketing dataset. The results of K nearest neighbours and Naïve Bayes are less while rest of the algorithms have given more or less same result with minor differences.As per the feature importance of XGBOOST it is clear that bank should focus more on clients with success in previous campaign. Whether client uses cellular phone or not and the month in which client is being called play a vital role in prediction. One thing should be noted that this modelling is based on behaviour of clients and not on their motivations. The features reveal the actions of client but not his/her thought process. So more descriptive features can be useful here for example interview summery. In that case natural language processing will give better results. In these times of crisis preserving the relationship with best customers is more crucial than ever. Using these results bank can specifically target clients and gain higher success in their endeavours. Saving a lot of time by not focusing on clients with less probability is yet another advantages of this project.

**REFERENCES**

1. Machine Learning Bank Marketing Data Set :

[https://drive.google.com/drive/folders/1TWqgR1YfJ\_0EUtvZTgccOoElZbpV-](https://drive.google.com/drive/folders/1TWqgR1YfJ_0EUtvZTgccOoElZbpV-QzG)

QzG

1. For Learning Python Language [Learn](https://www.learnpython.org/) [Python](https://www.learnpython.org/) [-](https://www.learnpython.org/) [Free](https://www.learnpython.org/) [Interactive](https://www.learnpython.org/) [Python](https://www.learnpython.org/) [Tutorial](https://www.learnpython.org/)

**APPENDIX**

A. SCREEN SHOTS



