Analysis of Machine Learning Models in Breast Cancer Detection

Project report submitted in partial fulfillment of the requirement for the degree of

BACHELOR OF TECHNOLOGY

IN

ELECTRONICS AND COMMUNICATION ENGINEERING

By

Vaibhav Rastogi

UNDER THE GUIDANCE OF

Dr. Nafis Uddin Khan

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY,
WAKNAGHAT
May 2022

TABLE OF CONTENTS

DECLARATION

We hereby declare that the work reported in the B.Tech Project Report entitled **"Analysis of Machine Learning Models in Breast Cancer Detection"** submitted at **Jaypee University of Information Technology, Waknaghat, India** is an authentic record of our work carried out under the supervision of **Dr. Nafis Uddin Khan.** We have not submitted this work elsewhere for any other degree or diploma.

Signature of Student:

Vaibhav Rastogi
181027

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Signature of the Supervisor:

Dr. Nafis Uddin Khan
Date:

Head of the Department/Project Coordinator

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to my mentor Dr. Naffis Uddin Khan who guided and supported me throughout the project "Analysis of Machine Learning Models in Breast Cancer Detection". The suggestions and instructions has served as a major contributor towards the completion of this project.

Also, I would like to thank my friends who provided me with innovative ideas and helped me in programming which helped me immensely in my project.

LIST OF FIGURES

ABSTRACT

This project focuses on detecting harmful cancerous cells in the body. Using classification model we can classify the cancerous cells as malignant or benign. Classification will be done based on the feature parameters provided in the dataset. This classification can help us determine cancerous cells in ones body accurately.The project uses different machine learning models and then compares the accuracy of all the models.

CHAPTER 1

INTRODUCTION

1.1. Introduction

Every year, Pathologists diagnose 14 million new patients with cancer around the world. That‟s millions of people who‟ll face years of uncertainty. Apart from conventional methods like mammogram, Machine Learning Algorithms are used to detect and predict the presence of cancerous cells in human body using machine learning techniques. Thus, the objective is to predict the presence of cancerous cells in the human body using Machine Learning Techniques, based on relative parameters. Machine learning is a branch of artificial intelligence that employs a variety of statistical, probabilistic and optimization techniques. It allows computers to "learn" from past examples and to detect hard-to-discern patterns from large, noisy or complex data sets.

1.2. What is Cancer?

Cancer- A word that shakes every person from top to bottom when heard. Cancer, considered a very dangerous illness can be a combination of emotional, physical, social, financial problems for an individual being. Cancer is a group of more than 100 diseases. Pathologists have been performing cancer diagnoses and prognoses for decades. Most pathologists have a 96–98% success rate for diagnosing cancer. They‟re pretty good at that part. The problem comes in the next part. According to the Oslo University Hospital, the accuracy of prognoses is only 60% for pathologists. We all know that cells are basic unit of life and they need to reproduce so that our body can use them.

Generally what happens is when cells get old, they die and are replaced by new cells. This is a normal process in which the human body functions. Cancer happens when there is a change in this cycle. A cancer patient has an abnormal process of cell reproduction. Sometimes, the cells start to grow uncontrollably and sometimes they don‟t. Cancer suffering patients also sometimes observe initial symptoms (different symptoms in different cancer types). The following are types of cancer:

1. Carcinoma- This is a type of cancer that develops in that part of the body which makes up the skin or the tissues. Example- livers or kidneys.

2. Sarcoma- This is the variant of cancer that begins in the bones and in soft tissues. This includes blood vessels, fat nerves etc.

3. Leukaemia- Blood or bone marrow cancer is known as Leukaemia. It normally affects the white blood cells or leukocytes.
4. Lymphoma- The immune system cancer is known as Lymphoma. It is the type of cancer that is developed in the immune system of the body.

1.3. Breast Cancer: An Overview

Breast cancer is a cancer that forms in the cells of the breasts. After skin cancer, breast cancer is the most common cancer diagnosed in women. It is one of the most common causes of death amongst women in the world. It alone is expected to account for 25% of all new cancer diagnoses and 15% of all cancer deaths among women worldwide. The causes of breast cancer can be increasing age, family history of breast cancer, radiation exposure, obesity, beginning menopause at older age etc. Other than these common causes, personal family history can also be one of the factors of suffering by breast cancer. In the early stages, particularly in case of breast cancer, it may not show any symptoms. But still, an abnormality may be observed on a mammogram. Depending on the type of breast cancer, the symptoms may vary, but the most common symptoms are:
1. Breast pain
2. Swelling in parts of breast
3. Discharge other than breast milk
4. Blood discharge

Bringing a change in lifestyle can be one of the ways to prevent breast cancer. Apart from this, exercise, proper medication, maintaining a healthy weight and choosing a healthy diet can be very helpful in prevention and maintaining.

1.4. What is Machine Learning?

Machine learning is a branch of artificial intelligence that employs a variety of statistical, probabilistic and optimization techniques. It allows computers to “learn” from past examples and to detect hard-to-discern patterns from large, noisy or complex data sets.

In 1959, Arthur Samuel, an American Pioneer in the field of computer gaming and artificial intelligence was the first one to term the coin “Machine Learning” Machine Learning (ML) is one of the core branches of Artificial Intelligence. It’s a system that takes in data, finds patterns, trains itself using the data and outputs an outcome. The main motive of machine learning is to let the computer learn from the given data automatically, without the interference of any external body.

The data in Machine Learning is divided into Test data, Training Data and Validation Data. Generally 20- 40% of the data is classified as test data and rest as training data. The training data is the part of data that we use to train the machine. The Validation data is that part which is used for frequent model evaluation and test data is that data which is used for the prediction of the result.

The general terminologies used in Machine Learning are:

1. Dataset - It is the set of all observations on which different techniques are applied as per the problem.

2. Feature- The Feature are the important factors of the data that help us to understand and visualize the problem. The system is fed with machine learning algorithms which in-turn are fed with the features so that the system can learn from them.

3. Model- It can be said that a Model is the representation of a phenomenon that a Machine Learning Algorithm has been learnt by the system. It learns from the data it is fed during the training. After learning from the data, the system implies the learnt techniques on the test data and the output that is presented is known as the Model.

CHAPTER 2

LITERATURE REVIEW

Machine learning is a field of Artificial Intelligence that allows a machine to automatically reason and learn from previous experiences. It helps the systems to reach to a conclusion while incorporating all the practical and emotional parameters. It allows the computer to learn from the examples and practices done earlier on them [1]. The branch is sub-divided into two types. Machine learning employs various algorithms depending on the type of problem and the dataset [2]. It has majorly been used as a practical application rather than theoretical and most widely used in the detection and prognosis of Cancer. Earlier tests like MRI, CT scan and X-rays were performed for the same motive, but the accuracy of these tests was very low and mainly, the indicative precision of a patient relies upon doctor's understanding, manifestations and affirmed analysis and still, at the end of the day, the outcomes can't be ensured [3]. So, ML was implemented in this field and it was seen that the accuracy of the results was much higher than ever before. After skin cancer, Breast cancer is the second most wide-spread cancer in the world and is considered deadly for women [3].Thus it was considered vital to control the widespread and the usage of ML became more and more popular [4]. On one hand, the conventional tests took days to give out the result and on the other hand, it was just a matter of few hours for ML to predict the cancer cells. Breast cancer is sub-divided into benign and malignant and the classification of the two is an important topic to research on [3]. So using machine learning, this process became simpler and reliable and in today‟s world, ML approaches are employed in modelling of cancer prognosis and progression. This study is based on genetic programming and machine learning algorithms that aim to construct a system to accurately differentiate between benign and malignant breast tumours. The aim of this study was to optimize the learning algorithm [8]. The paper proposes a hybrid model combined of several Machine Learning (ML) algorithms including Support Vector Machine (SVM), Artificial Neural Network (ANN), K-Nearest Neighbor (KNN) and Decision Tree for effective breast cancer detection . This study also discusses the datasets used for breast cancer detection and diagnosis. The proposed model can be used with different data types such as image, blood, etc [9]. Principal component analysis is the oldest and best known technique of multivariate data analysis. It was first coined by Pearson (1901), and developed independently by Hotelling (1933). Like many other multivariate methods, it was not widely accepted nor used until the advent of electronic computers, but it is now well entrenched in virtually every statistical software packages [5]. Principal Component Analysis (PCA) is the general name for a technique which uses sophisticated underlying mathematical principles to transforms a number of possibly correlated variables into a smaller number of variables called principal components. PCA is a multi-variable technique that analyses a data table in which observations can be described by many inter-correlated quantitative dependent variables. Its goal is to extract the important information from the statistical data to represent it as a set of new orthogonal variables called principal components, and to display the pattern of similarity between the observations and of the variables as points in spot maps [6]. Many information processing problems can be transformed into some form of Eigen value or singular value problems. Eigen value decomposition (EVD) and singular value decomposition (SVD) are usually used for solving these problems. PCA is a statistical method that is directly related to EVD and SVD. Unsupervised learning techniques with feature extraction utilising various neural network models are used. Orthogonal decomposition is a well-known approach for eliminating ill-conditioning.

CHAPTER 3

SOFTWARE/SIMULATION

3.1. Block Diagram

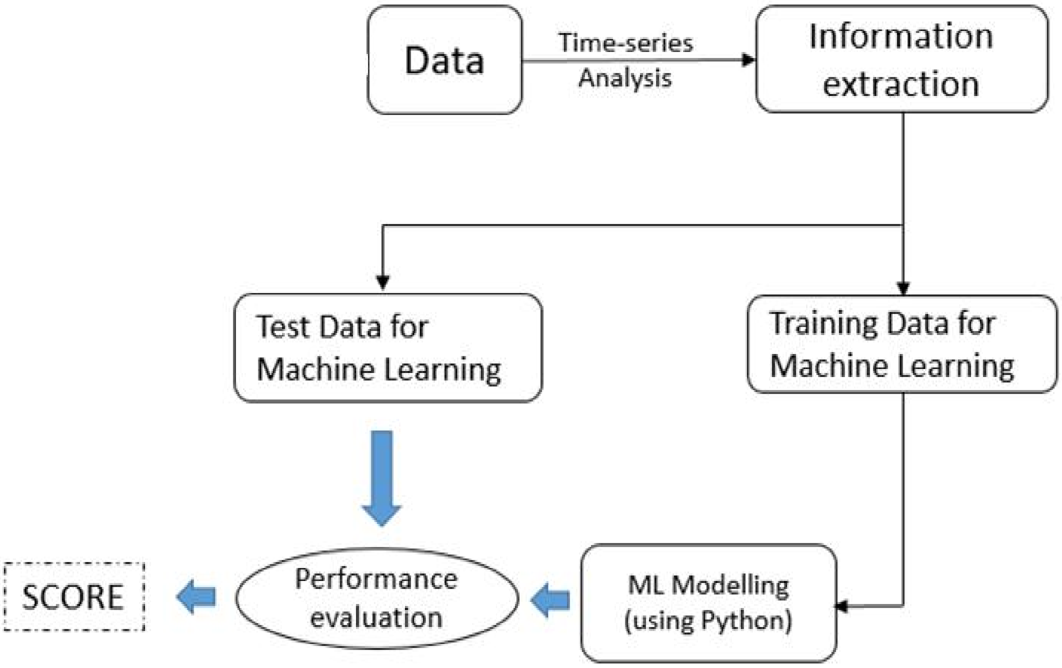

Data
Time-series
Analysis
Information extraction
Test Data for Machine Learning
Training Data for Machine Learning
SCORE
Performance evaluation
ML Modelling (using Python)

Figure 3.1: Block Diagram

The software that is used to implement Machine Learning in our project is Anaconda. Anaconda is an open source, free platform for python data science, for machine learning. Anaconda Navigator is a graphical user interface that may be used instead of the command line interface for graphs. The Anaconda Navigator is a desktop GUI that includes many packages. It can also be used to search for packages on anaconda cloud. Some of the default applications of Navigator are:

1. Jupyter Lab
2. Jupyter Notebook
3. Spyder
4. Glueviz
5. RStudio

The application used in our project is Jupyter IDE. Jupyter is a Python programming environment for scientists (a free integrated development environment- IDE). It has a unique combination and offers analysis, debugging and editing tool for database.

3.2. Types of Machine Learning:

There are basically 3 broad categories of Machine Learning, depending upon the nature of the learning, available to the learning system:

1. Supervised Learning: A supervised learning algorithm is an algorithm which learns by the given inputs and its corresponding outputs of the given data. Basically, we need to train the machine according to the labelled columns and rows. This type of learning is exactly like a student learning under the supervision of his teacher. It is considered helpful since the teacher will help the student to understand by giving relative examples. In supervised machine learning, we have an input variable “x” and an output variable “y=f(x)”, then we use an algorithm to map the input to the output. The techniques used in Supervised Machine Learning are Regression and Classification.

a. Regression: The process of developing a model or function for converting data into continuous real values rather than utilising classes or discrete values is known as regression. Based on previous data, it may also determine dispersion movement.The skill of a regression predictive model must be expressed as an error in those predictions since it predicts a quantity.

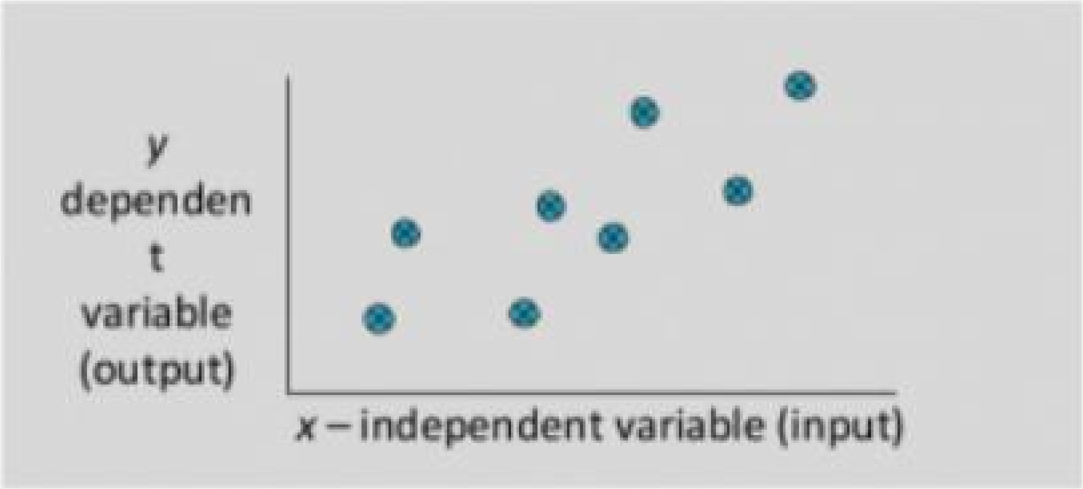


Figure 3.2.1: Distribution movement in Regression

Terminologies used in regression:

I. Outliers- Assume there is a dataset observation with a high or low value an incentive when contrasted with different observations in the given data, for instance, it does not have a position in the population; this type of observation is known as an outlier. In straightforward words, it is the extreme value. It is basically an issue on the grounds that multiple occasions it hampers the outcomes we get.

II. Multicollinearity-At the point when the non-dependent variables are profoundly related to one another then they are considered to be Multicollinear.Multicollinearity should not be present in the dataset, according to a variety of regression methodologies. It makes selecting the most important independent variable more difficult (factor).

III. Heteroscedasticity- When variability of dependent variable is not equal across of an independent variable, this case is known as Heteroscedasticity. For ex- As one's income increments, one ought to spend more money on ourselves to look better, whereas a poor person spends constant amount of money. People with high incomes display a greater variability of spent money

.IV. Underfitting and Overfitting: Overfitting is the case in which the algorithm shows great results on the training set but underperforms in the case of test set. It's also called problem of high variance. Underfitting is the case in which the algorithm is so poor that it barely fits into the training set. Its known as the problem of high bias Depending on the type of input and type of output required, the following are types of regressions and they differ from each other in terms of independent and dependent variables:

i. Linear Regression- In this training model the dependent variable,s nature is continuous. Both variables are related to each other in Linear nature. The equation of Linear regression is:

$$Y = \beta_o + \beta_1 Xi + \in i$$

Where,

Yi is the Dependent Variable

βo is the Y intercept

β1Xi is the Slope coefficient and the Independent Variable respectively and

εi is the Random Error

ii. Polynomial Regression- Polynomial regression- It is basically a technique used in the case of non-linear functions. In this, the dependent and independent variables are related to each other in a Non-Linear relation. Polynomial Regression's general equation is:

$$Y= \theta o + \theta_1 X + \theta_2 X^2 + \ldots + \theta_m X^m + \text{residual error}$$

iii. Logistic Regression-The dependent variable is divided into two categories or is binary in nature, whilst the independent variables are either continuous or binary. The general equation of logistic regression is:

$$p = \frac{1}{1 + e^{-(bo + b_1X_1 + b_2X_2 + \cdots + b_kX_k)}}$$

iv. Quantile Regression- This is basically an extended version of Linear Regression. When the data contains outliers, high skewness, and heteroscedasticity, this model is applied.

v. Ridge Regression- The Ridge Regression is a technique used when there exists multicollinearity in the data. It is used for continuous value prediction problem.

vi. Lasso Regression- Lasso is abbrevation for Least Absolute Shrinkage and Selection Operator. It uses the concept of Shrinkage. It means the data values are centred towards the centre or the mean. It is generally used in models with less number of parameters or features.

vii. Elastic Net Regression- This technique is preferred over the lasso and the ridge regression. It combines the process of feature elimination and feature coefficient reduction from the model to improve the predictions.

viii. Principle Components Regression: It is used when there are many independent variables or multicollinearity existing in the data. Firstly, it gets the principal components then run regression analysis on principal components. Dimensionality reduction and removal of multicollinearity are its key features.

ix. Partial Least Squares Regression: It is used as an alternative to principal component regression when the independent variables are highly linked.

x. Support Vector Regression: It can provide us with the solution of both linear as well as non linear models. SVM employs non-linear kernel functions to identify the best solution for non-linear models (such as polynomial).

xi. Ordinal Regression: This model is used for prediction of ranked values. When the dependent variable is ordinal, this method of regression is appropriate.

xii. Poisson Regression: Poisson regression is used when the dependent variable has count data. The dependent variable must have a Poisson distribution whose counts cannot be negative, thus making it unfit for non-whole numbers.

xiii. Negative Binomial Regression:It works with count data as well as Poisson Regression. "How is it different from poisson regression?" one could wonder. The variance of a count distribution is not assumed to be equal to its mean in negative binomial regression. When using poisson regression, the variance is assumed to be equal to the mean.

xiv. Quassi Poisson Regression: It's a different approach to negative binomial regression. It may also be applied to count data that is distributed. A quasi-Poisson model's variance is a linear function of the mean, whereas a negative binomial model's variance is a quadratic function of the mean.

xv. Cox Regression: It is suitable for time-to-event data as well as estimating the time it takes to reach a certain event, survival analysis can also be used to compare time-to-event for multiple groups.

xvi. Torbit Regression: When the dependent variable has censoring, it is used to calculate linear relationships between variables. When we witness the independent variable for all observations, but only know the real value of the dependent variable for a limited range of observations, this is referred to as censoring.

b. Classification- The process of identifying or inventing a model or function that aids in the separation of data into various categorical classes (i.e. discrete values) is known as classification. In classification, data is classified into different labels based on input characteristics, and the labels are then predicted for the data. The result of classification is a category, such as whether a human being is male or female.

The types of algorithms in classification are:

1. Linear Classifiers- Logistic Regression and Naïve Byes Classifier
2. Nearest Neighbour
3. Support Vector Machines
4. Decision Trees
5. Boosted Trees
6. Random Forest
7. Neural Networks

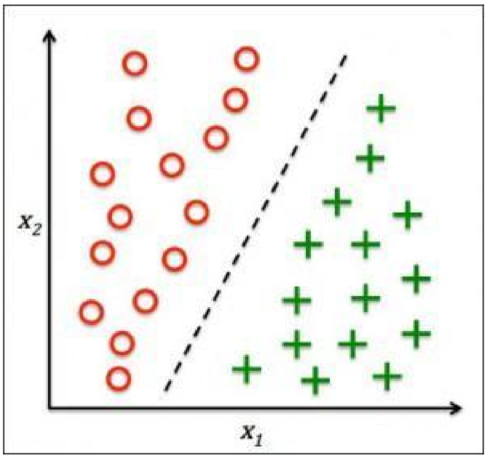


Figure 3.2.2: Classification of Supervised Data

2. Unsupervised Learning: The machine is not given any sort of guidance. The machine itself learns from the pattern or program and acts on the result accordingly. This type of learning may be considered helpful because if we ourselves predict the output, it may be of some kind other than the outcome predicted by the machine. This type of learning helps to get different outcomes of same kind since the machine may look at the input from a different point of view than ours. Unsupervised learning is classified as follows:

i. Clustering- Clustering is done basically to categorize the data in form of groups. It is concerned with identifying patterns in uncategorized data sets. The machine will process the data and form clusters of the data as the final output. Clustering is of many types:

a) Hierarchical Clustering

b) K-means cluster

c) K-Nearest Neighbour

d) Principle Component Analysis

e) Singular Value Decomposition

f) Independent Component Analysis

ii. Association: Association is all about associating data variables (objects) in large databases. For example- Rating of a movie, based on reviews of different people.

3. Reinforcement Learning: Using this algorithm, the machine is taught to make particular judgments. Working of this algorithm is as follows: the machine is placed in a situation where it must continually learn by trial and error. This system learns from its failures and tries to collect as much data as possible in order to get the best conclusions possible.

3.3. Breast Cancer

Breast Cancer is classified into the following two categories:

a) Malignant- This type of cancerous cells are those cells that spread from one cell to another

a) Benign- These types of cancerous cells are those that doesn"t spread from one cell to another but they can be severe if they are present in blood vessels or nerves since they can cause blockage.

3.4. Data Set

The data used in the project was taken from the home page for the University of California. The database has the following URL:

http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28diagnostic%29

The dataset consists of 569 columns and 32 rows. Columns consist of different person and rows consist of different parameters of breast cancer. The parameters are as follows:

1. Diagnosis-Malignant or Benign
2. Radius Mean
3. Texture Mean
4. Perimeter Mean
5. Area Mean
6. Smoothness Mean
7. Compactness Mean
8. Concavity Mean
9. Concave points Mean
10. Symmetry Mean
11. Fractal Dimension Mean
12. Radius se
13. Texture se
14. Perimeter se
15. Area se
16. Smoothness se
17. Compactness se
18. Concavity se
19. Concave Points se
20. Symmetry se
21. Fractal Dimensions se
22. Radius worst
23. Texture worst
24. Perimeter worst
25. Area worst
26. Smoothness worst
27. Compactness worst
28. Concavity worst
29. Concave points worst

30. Symmetry worst

31. Fractal Dimensions worst

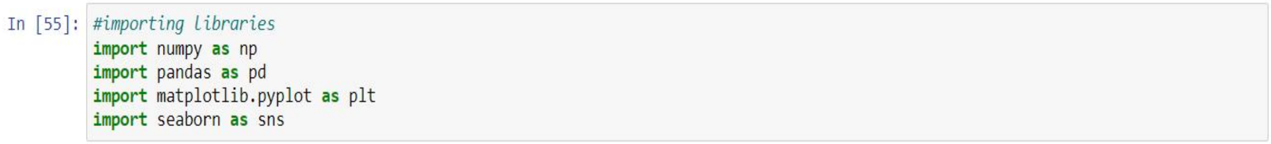
3.5. The Code

Getting started with the code, the project consists of various steps that were used to compile the code which are as follows:

1. IMPORTING THE PYTHON LIBRARIES

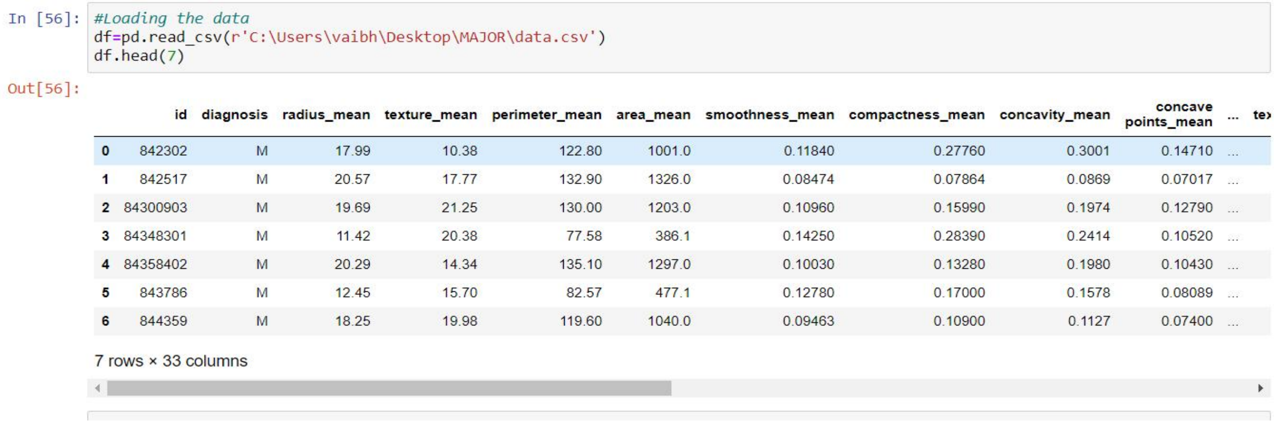
The libraries used in the code are:

1. NumPy
2. Pandas
3. MatpotLib
4. Seaborn



```
In [55]: #importing libraries
         import numpy as np
         import pandas as pd
         import matplotlib.pyplot as plt
         import seaborn as sns
```

2. LOADING THE DATASET



```
In [56]: #Loading the data
         df=pd.read_csv(r'C:\Users\vaibh\Desktop\MAJOR\data.csv')
         df.head(7)
```

Out[56]:

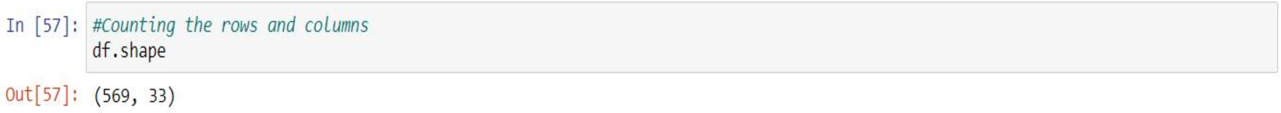
	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	tex
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...	
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...	
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...	
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...	
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...	
5	843786	M	12.45	15.70	82.57	477.1	0.12780	0.17000	0.1578	0.08089	...	
6	844359	M	18.25	19.98	119.60	1040.0	0.09463	0.10900	0.1127	0.07400	...	

7 rows × 33 columns

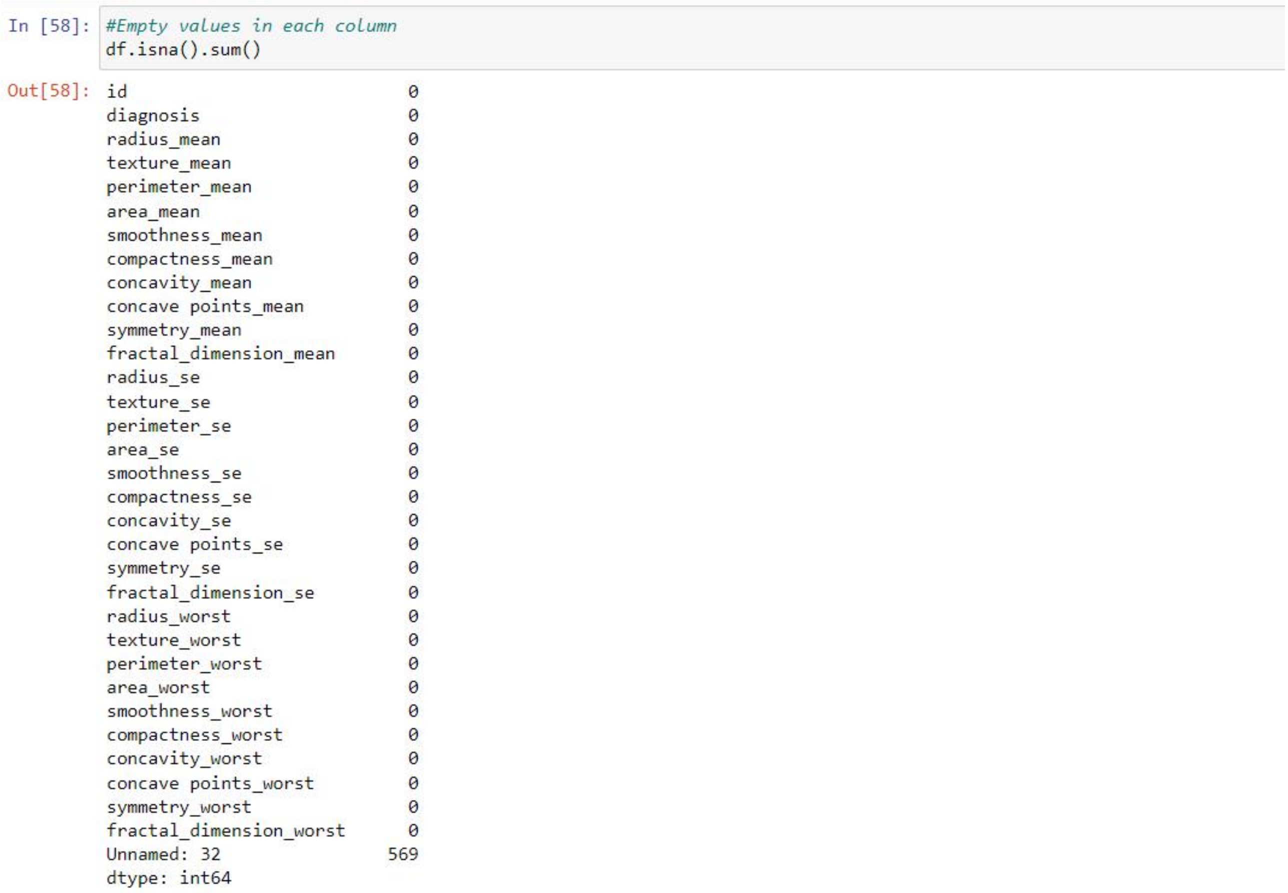
The dataset is uploaded and first 7 rows are printed.

3. PROCESSING THE DATA

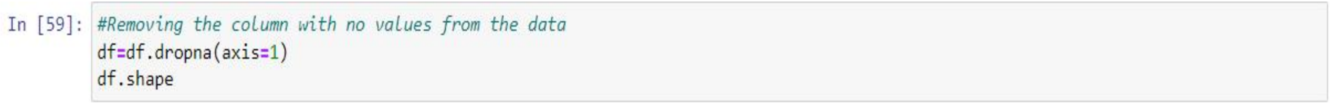
The data is then compiled to check for its size and find out if there are any empty values such as NaN, na, NAN.



```
In [57]: #Counting the rows and columns
         df.shape

Out[57]: (569, 33)
```



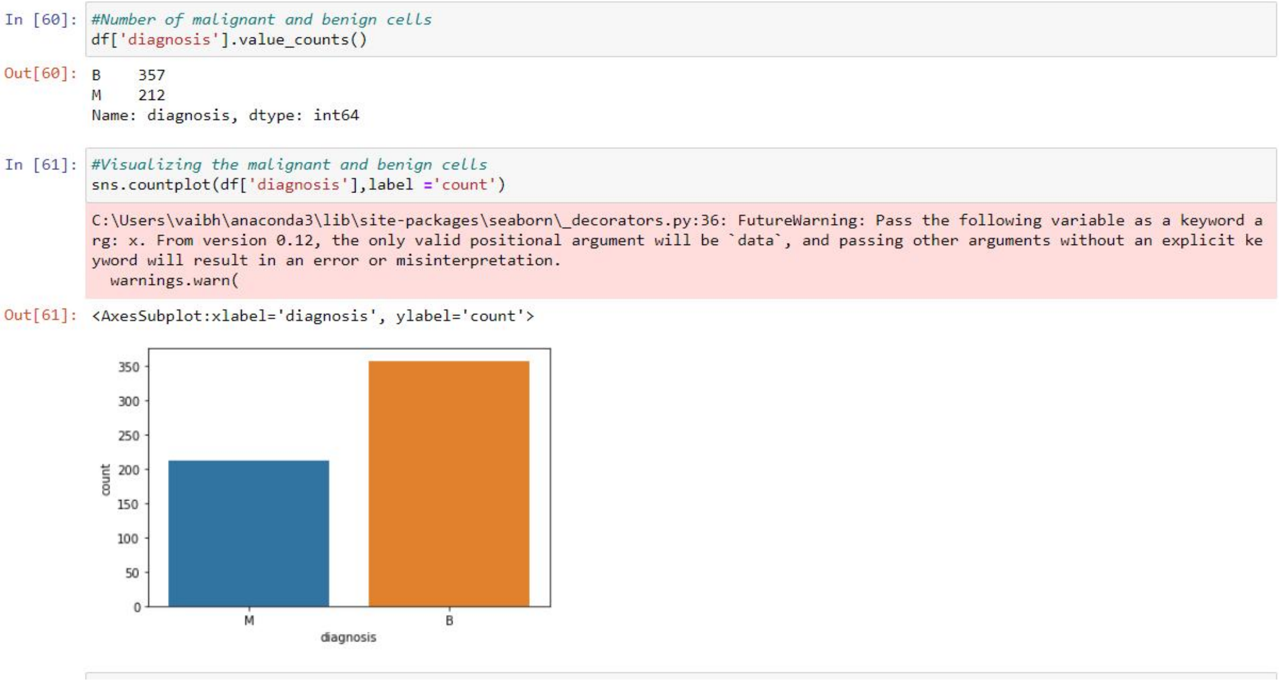
```
In [58]: #Empty values in each column
         df.isna().sum()

Out[58]: id                           0
         diagnosis                    0
         radius_mean                  0
         texture_mean                 0
         perimeter_mean               0
         area_mean                    0
         smoothness_mean              0
         compactness_mean             0
         concavity_mean               0
         concave points_mean          0
         symmetry_mean                0
         fractal_dimension_mean       0
         radius_se                    0
         texture_se                   0
         perimeter_se                 0
         area_se                      0
         smoothness_se                0
         compactness_se               0
         concavity_se                 0
         concave points_se            0
         symmetry_se                  0
         fractal_dimension_se         0
         radius_worst                 0
         texture_worst                0
         perimeter_worst              0
         area_worst                   0
         smoothness_worst             0
         compactness_worst            0
         concavity_worst              0
         concave points_worst         0
         symmetry_worst               0
         fractal_dimension_worst      0
         Unnamed: 32                569
         dtype: int64
```



```
In [59]: #Removing the column with no values from the data
         df=df.dropna(axis=1)
         df.shape
```

4. SELECTING A PARAMETER

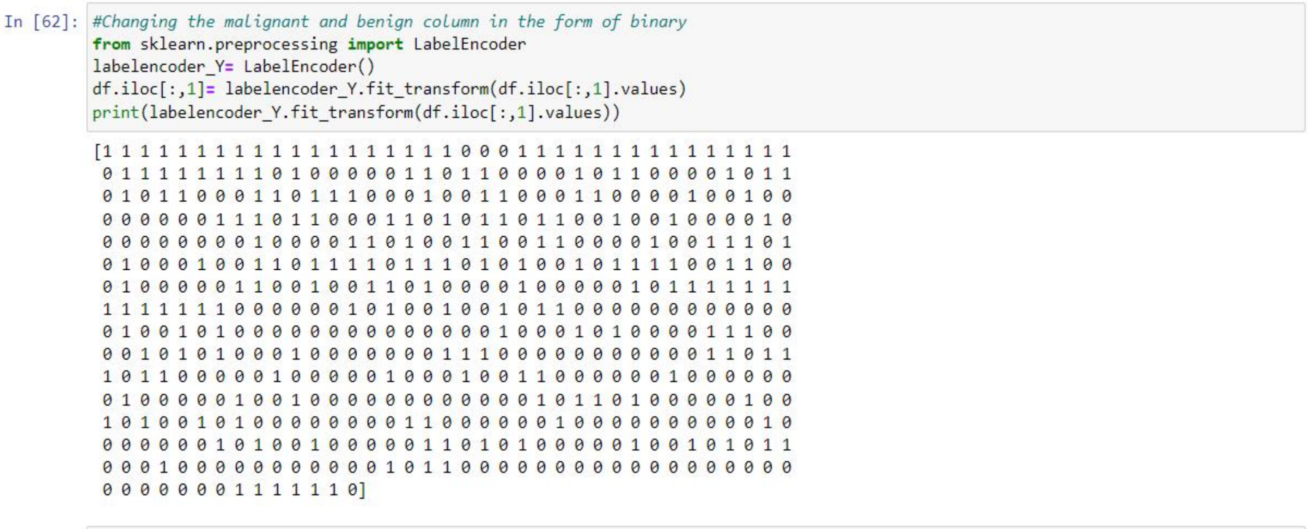
A label/feature is selected to implement Machine Learning algorithm upon which in this case is “diagnosis” which consists of inputs as „M‟ for malignant and „B‟ for benign. Then we count the number of malignant and benign data to work upon.



```
In [60]: #Number of malignant and benign cells
         df['diagnosis'].value_counts()

Out[60]: B    357
         M    212
         Name: diagnosis, dtype: int64

In [61]: #Visualizing the malignant and benign cells
         sns.countplot(df['diagnosis'],label ='count')

         C:\Users\vaibh\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variable as a keyword a
         rg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit ke
         yword will result in an error or misinterpretation.
           warnings.warn(

Out[61]: <AxesSubplot:xlabel='diagnosis', ylabel='count'>
```

5. REPLACING THE STRINGS WITH INTEGERS

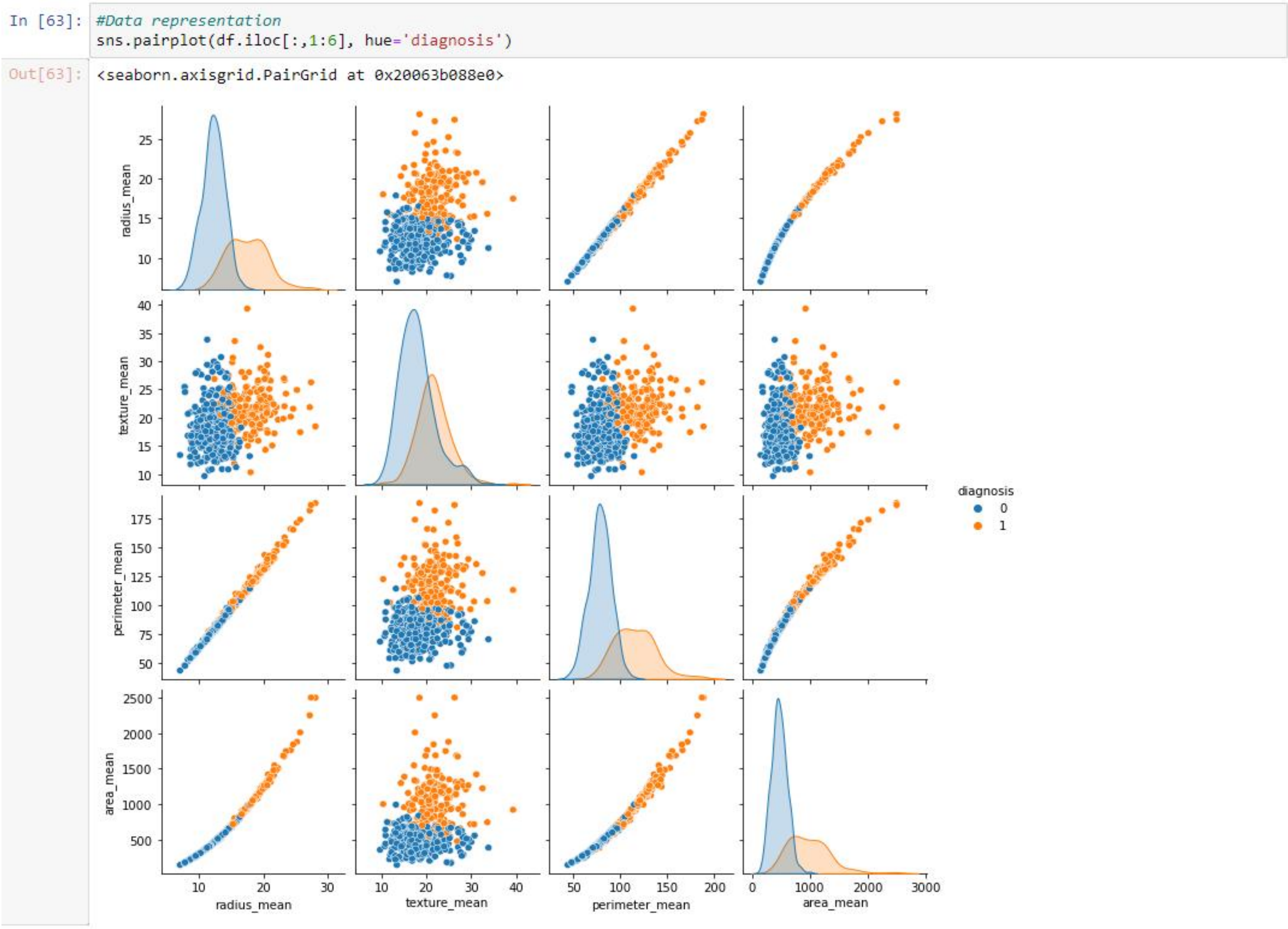
The input of “M‟ and “B‟ is replaced with “1‟ and “0‟.



```
In [62]: #Changing the malignant and benign column in the form of binary
         from sklearn.preprocessing import LabelEncoder
         labelencoder_Y= LabelEncoder()
         df.iloc[:,1]= labelencoder_Y.fit_transform(df.iloc[:,1].values)
         print(labelencoder_Y.fit_transform(df.iloc[:,1].values))

         [1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
          0 1 1 1 1 1 1 1 1 0 1 0 0 0 0 0 1 1 0 1 1 0 0 0 0 1 0 1 1 0 0 0 0 1 0 1 1
          0 1 0 1 1 0 0 0 1 1 0 1 1 1 0 0 0 1 0 0 1 1 0 0 0 1 1 0 0 0 0 1 0 0 1 0 0
          0 0 0 0 0 0 1 1 1 0 1 1 0 0 0 1 1 0 1 0 1 1 0 1 1 0 0 1 0 0 1 0 0 0 0 1 0
          0 0 0 0 0 0 0 0 1 0 0 0 0 1 1 0 1 0 0 1 1 0 0 1 1 0 0 0 0 1 0 0 1 1 1 0 1
          0 1 0 0 0 1 0 0 1 1 0 1 1 1 1 0 1 1 1 0 1 0 1 0 0 1 0 1 1 1 1 0 0 1 1 0 0
          0 1 0 0 0 0 0 1 1 0 0 1 0 0 1 1 0 1 0 0 0 0 1 0 0 0 0 0 1 0 1 1 1 1 1 1 1
          1 1 1 1 1 1 1 0 0 0 0 0 0 1 0 1 0 0 1 0 0 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0
          0 1 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 1 0 0 0 0 1 1 1 0 0
          0 0 1 0 1 0 1 0 0 0 1 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1
          1 0 1 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 1 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0
          0 1 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 0 1 0 0 0 0 0 1 0 0
          1 0 1 0 0 1 0 1 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0
          0 0 0 0 0 0 1 0 1 0 0 1 0 0 0 0 0 1 1 0 1 0 1 0 0 0 0 0 1 0 0 1 0 1 0 1 1
          0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
          0 0 0 0 0 0 0 1 1 1 1 1 1 0]
```

5. PAIR PLOTTING

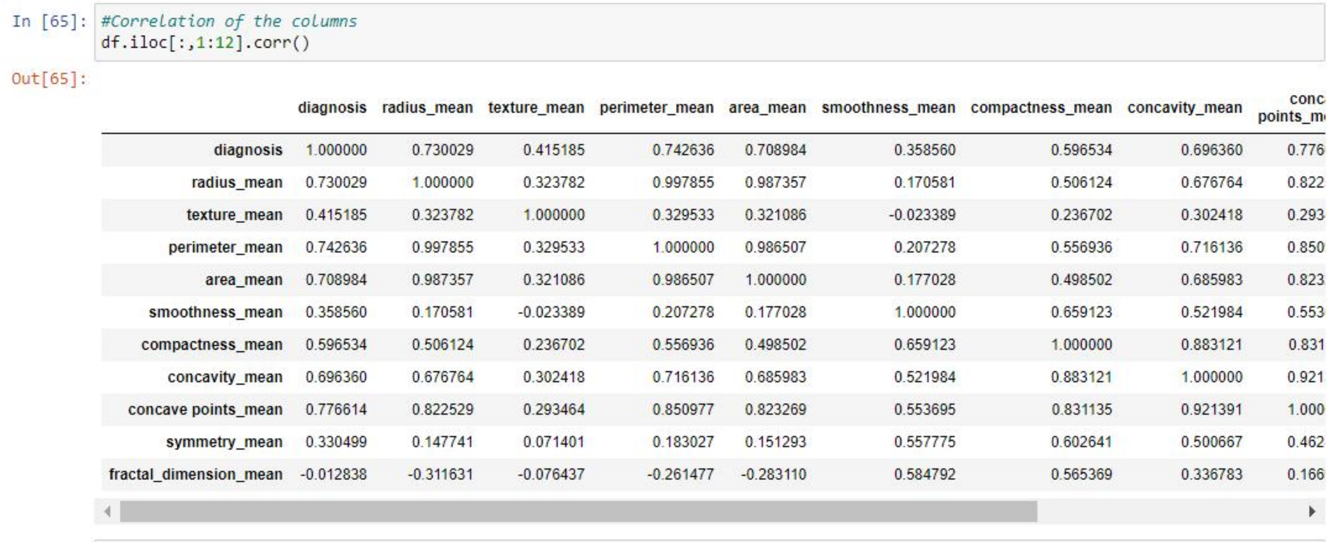
A "pairs plot," sometimes known as a scatter plot, is a graph in which the value of one variable in the same data row is compared with the value of another variable.



```
In [63]: #Data representation
         sns.pairplot(df.iloc[:,1:6], hue='diagnosis')

Out[63]: <seaborn.axisgrid.PairGrid at 0x20063b088e0>
```

7. GETTING CORRELATION OF THE LABELS

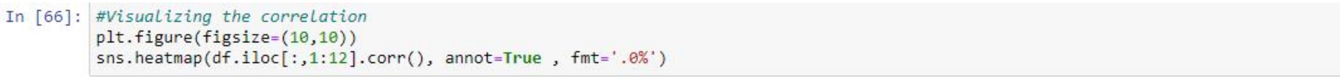
The columns are correlated with each other forming another table to provide the dependency of all columns on each other.



```
In [65]: #Correlation of the columns
         df.iloc[:,1:12].corr()

Out[65]:
```

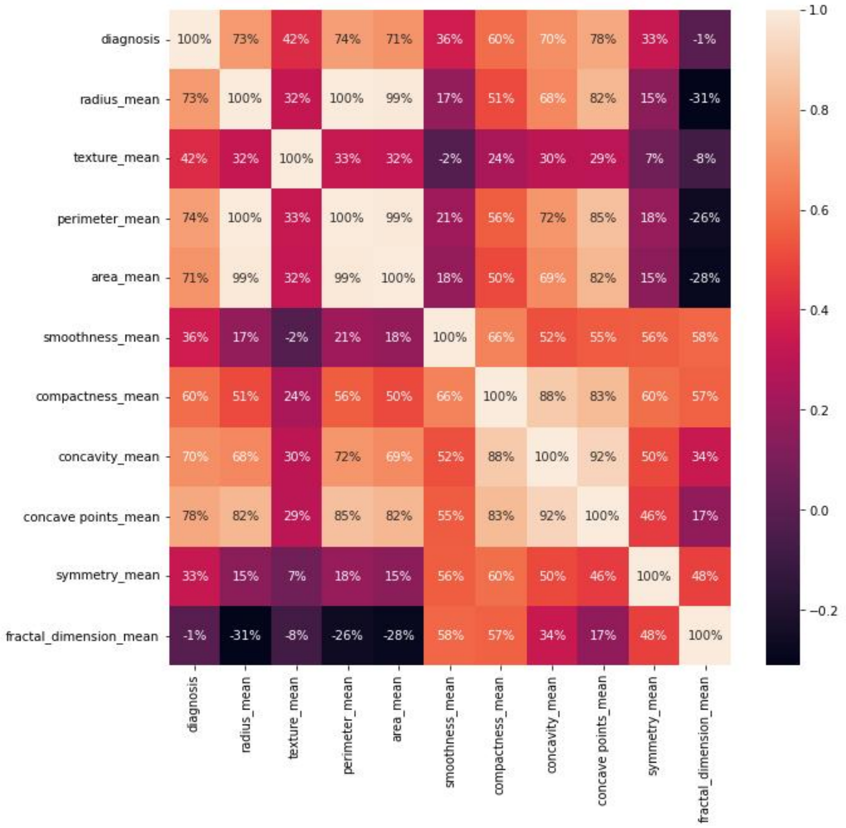
	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	conc points_m
diagnosis	1.000000	0.730029	0.415185	0.742636	0.708984	0.358560	0.596534	0.696360	0.776
radius_mean	0.730029	1.000000	0.323782	0.997855	0.987357	0.170581	0.506124	0.676764	0.822
texture_mean	0.415185	0.323782	1.000000	0.329533	0.321086	-0.023389	0.236702	0.302418	0.293
perimeter_mean	0.742636	0.997855	0.329533	1.000000	0.986507	0.207278	0.556936	0.716136	0.850
area_mean	0.708984	0.987357	0.321086	0.986507	1.000000	0.177028	0.498502	0.685983	0.823
smoothness_mean	0.358560	0.170581	-0.023389	0.207278	0.177028	1.000000	0.659123	0.521984	0.553
compactness_mean	0.596534	0.506124	0.236702	0.556936	0.498502	0.659123	1.000000	0.883121	0.831
concavity_mean	0.696360	0.676764	0.302418	0.716136	0.685983	0.521984	0.883121	1.000000	0.921
concave points_mean	0.776614	0.822529	0.293464	0.850977	0.823269	0.553695	0.831135	0.921391	1.000
symmetry_mean	0.330499	0.147741	0.071401	0.183027	0.151293	0.557775	0.602641	0.500667	0.462
fractal_dimension_mean	-0.012838	-0.311631	-0.076437	-0.261477	-0.283110	0.584792	0.565369	0.336783	0.166

8. REPRESENTING THE CORRELATION IN HEATMAP FORM



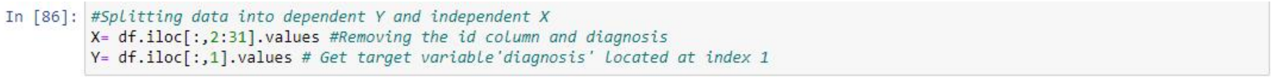
```
In [66]: #Visualizing the correlation
         plt.figure(figsize=(10,10))
         sns.heatmap(df.iloc[:,1:12].corr(), annot=True , fmt='.0%')
```

```
Out[66]: <AxesSubplot:>
```

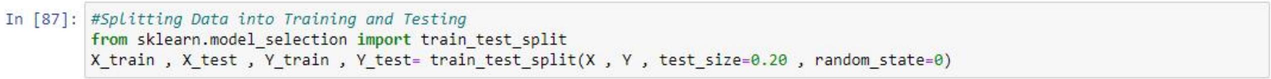


9. SPLITTING DATA

Firstly variables are separated into independent and dependent variables “X‟ being the independent ones containing all other features on which the „diagnosis‟ is dependent which is the dependent variable “Y‟.



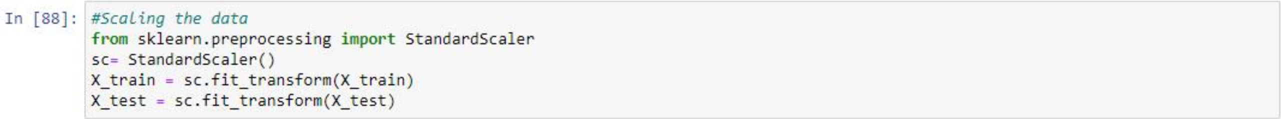
```
In [86]: #Splitting data into dependent Y and independent X
         X= df.iloc[:,2:31].values #Removing the id column and diagnosis
         Y= df.iloc[:,1].values # Get target variable'diagnosis' located at index 1
```



```
In [87]: #Splitting Data into Training and Testing
         from sklearn.model_selection import train_test_split
         X_train , X_test , Y_train , Y_test= train_test_split(X , Y , test_size=0.20 , random_state=0)
```

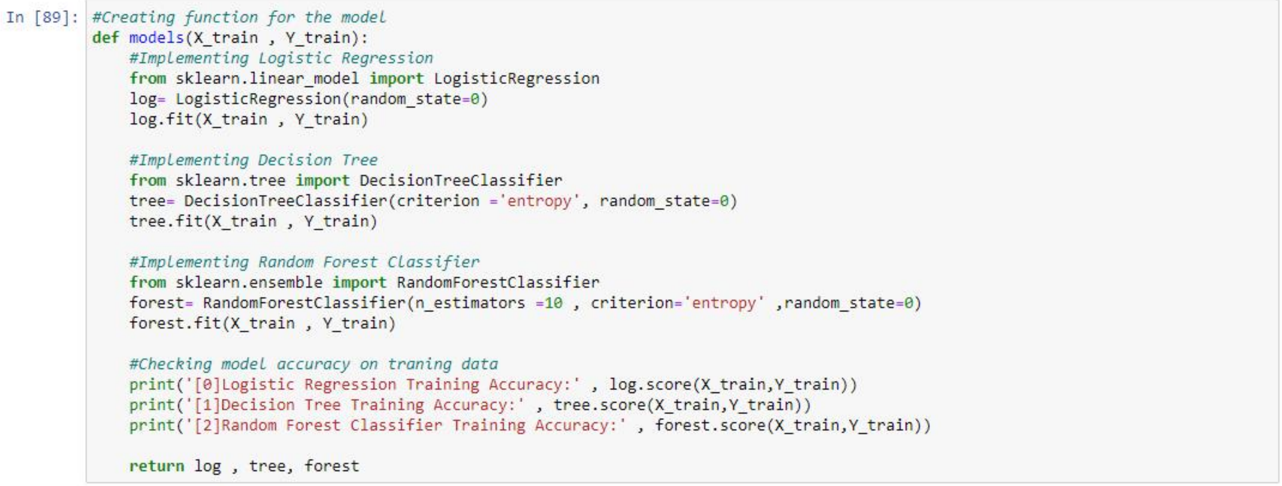
10. FEATURE SCALING

This means the data will be within a specific range for example zero to hundred or zero to one. It is the scaling of the data to bring all characteristics to the same magnitude.



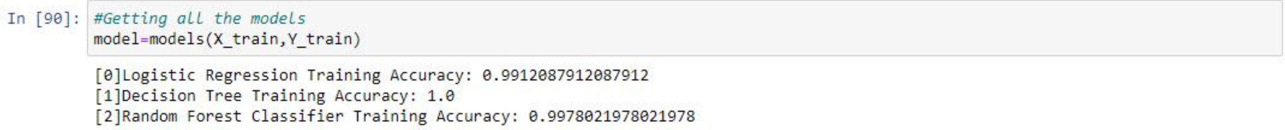
```
In [88]: #Scaling the data
         from sklearn.preprocessing import StandardScaler
         sc= StandardScaler()
         X_train = sc.fit_transform(X_train)
         X_test = sc.fit_transform(X_test)
```

11. IMPLEMENTING VARIOUS ML CLASSIFIERS/MODELS



```
In [89]: #Creating function for the model
         def models(X_train , Y_train):
             #Implementing Logistic Regression
             from sklearn.linear_model import LogisticRegression
             log= LogisticRegression(random_state=0)
             log.fit(X_train , Y_train)

             #Implementing Decision Tree
             from sklearn.tree import DecisionTreeClassifier
             tree= DecisionTreeClassifier(criterion ='entropy', random_state=0)
             tree.fit(X_train , Y_train)

             #Implementing Random Forest Classifier
             from sklearn.ensemble import RandomForestClassifier
             forest= RandomForestClassifier(n_estimators =10 , criterion='entropy' ,random_state=0)
             forest.fit(X_train , Y_train)

             #Checking model accuracy on traning data
             print('[0]Logistic Regression Training Accuracy:' , log.score(X_train,Y_train))
             print('[1]Decision Tree Training Accuracy:' , tree.score(X_train,Y_train))
             print('[2]Random Forest Classifier Training Accuracy:' , forest.score(X_train,Y_train))

             return log , tree, forest
```

The techniques are implemented on the training dataset to train it for the test dataset.

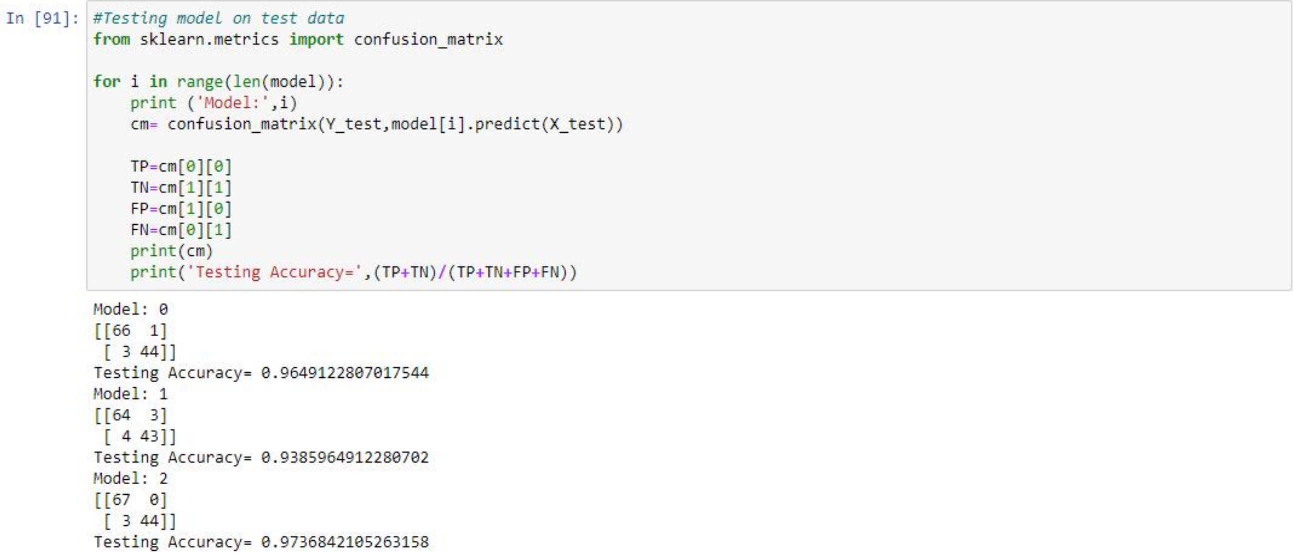


```
In [90]: #Getting all the models
         model=models(X_train,Y_train)

         [0]Logistic Regression Training Accuracy: 0.9912087912087912
         [1]Decision Tree Training Accuracy: 1.0
         [2]Random Forest Classifier Training Accuracy: 0.9978021978021978
```

Here the Decision Tree Classifier shows the best training accuracy of 100%.

12. CONFUSION MATRIX

A classification model's error matrix is a table that illustrates how well it works. It shows the ways in which classifier is confused in making predictions. Classification accuracy is measured on the basis of confusion matrix on the test data.

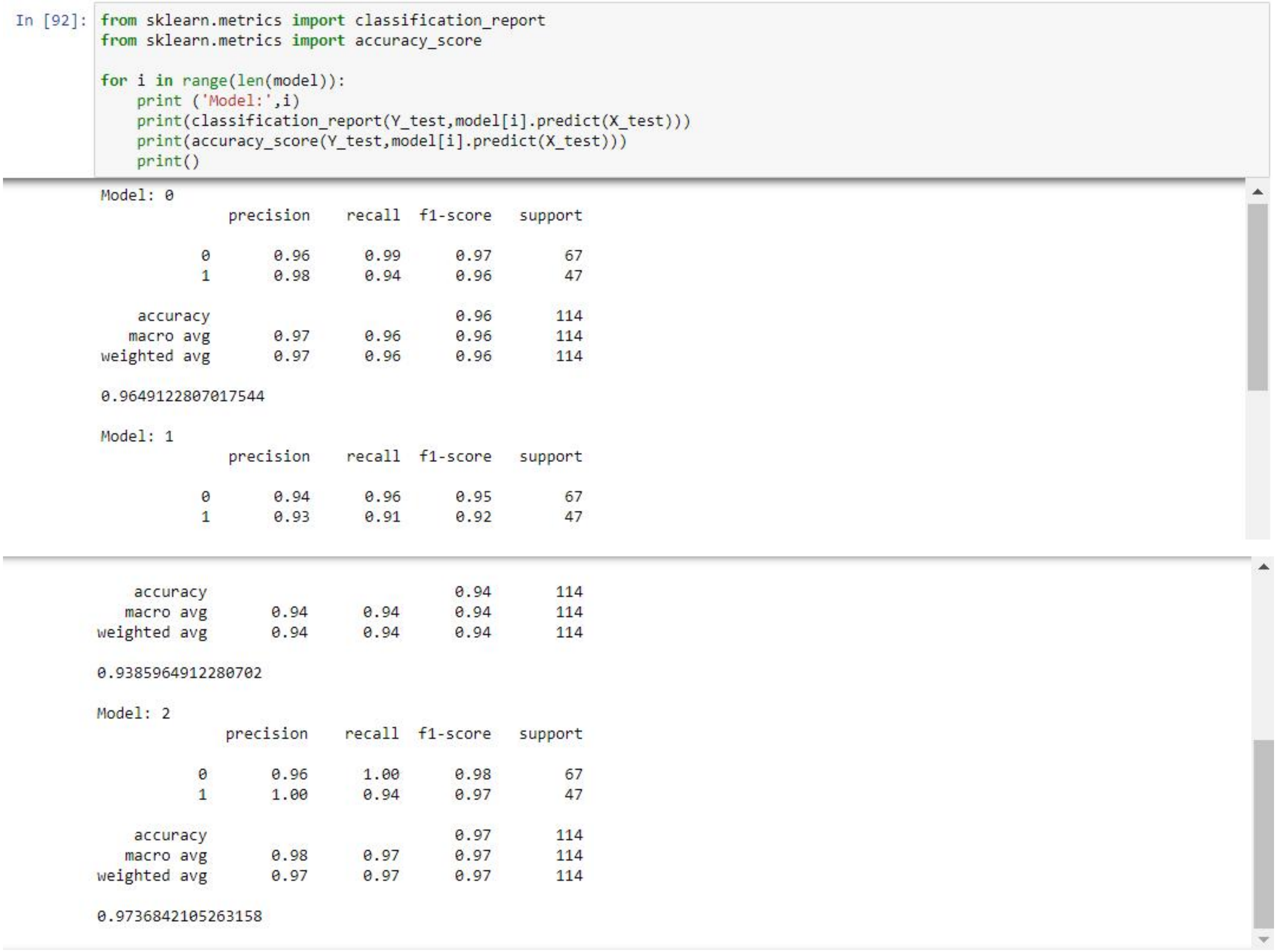


```
In [91]: #Testing model on test data
         from sklearn.metrics import confusion_matrix

         for i in range(len(model)):
             print ('Model:',i)
             cm= confusion_matrix(Y_test,model[i].predict(X_test))

             TP=cm[0][0]
             TN=cm[1][1]
             FP=cm[1][0]
             FN=cm[0][1]
             print(cm)
             print('Testing Accuracy=',(TP+TN)/(TP+TN+FP+FN))

         Model: 0
         [[66  1]
          [ 3 44]]
         Testing Accuracy= 0.9649122807017544
         Model: 1
         [[64  3]
          [ 4 43]]
         Testing Accuracy= 0.9385964912280702
         Model: 2
         [[67  0]
          [ 3 44]]
         Testing Accuracy= 0.9736842105263158
```

From these accuracy readings it can be said that the Logistic Regression and Random Forest Classifiers have the maximum accuracy of around 96.5% and 97.3% respectively.

13. THER METRICS APPLIED FOR MORE ACCURACY

Other parameters like precision, recall, and F1-score are calculated on the basis of confusion matrix on the test data.



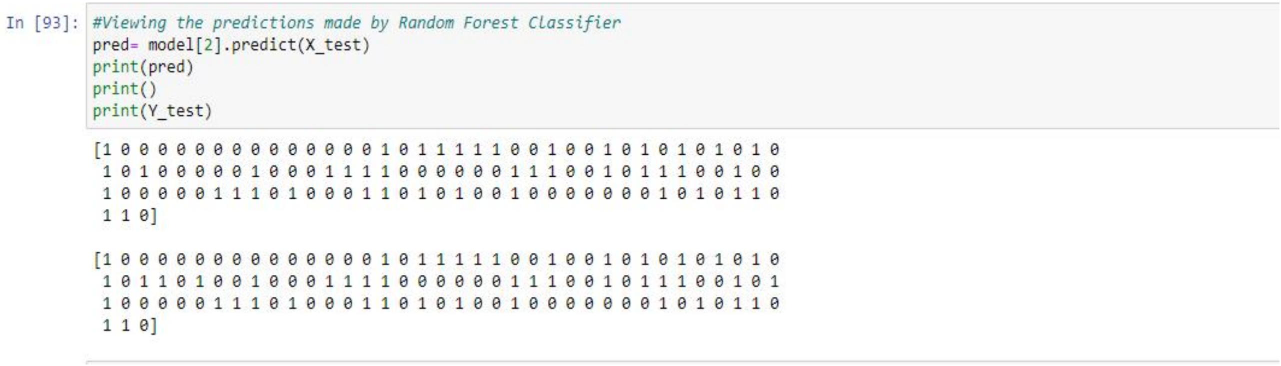
```
In [92]: from sklearn.metrics import classification_report
         from sklearn.metrics import accuracy_score

         for i in range(len(model)):
             print ('Model:',i)
             print(classification_report(Y_test,model[i].predict(X_test)))
             print(accuracy_score(Y_test,model[i].predict(X_test)))
             print()
```

```
Model: 0
              precision    recall  f1-score   support

           0       0.96      0.99      0.97        67
           1       0.98      0.94      0.96        47

    accuracy                           0.96       114
   macro avg       0.97      0.96      0.96       114
weighted avg       0.97      0.96      0.96       114

0.9649122807017544

Model: 1
              precision    recall  f1-score   support

           0       0.94      0.96      0.95        67
           1       0.93      0.91      0.92        47
```

```
    accuracy                           0.94       114
   macro avg       0.94      0.94      0.94       114
weighted avg       0.94      0.94      0.94       114

0.9385964912280702

Model: 2
              precision    recall  f1-score   support

           0       0.96      1.00      0.98        67
           1       1.00      0.94      0.97        47

    accuracy                           0.97       114
   macro avg       0.98      0.97      0.97       114
weighted avg       0.97      0.97      0.97       114

0.9736842105263158
```

14. SELECTION OF APT MACHINE LEARNING ALGORITHM

Based on the above calculations, Random Forest Classifier have highest accuracy so, Random Forest Classifier is selected to print the predicted data as well as the test dataset to be compared.



```
In [93]: #Viewing the predictions made by Random Forest Classifier
         pred= model[2].predict(X_test)
         print(pred)
         print()
         print(Y_test)

         [1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 1 1 1 0 0 1 0 0 1 0 1 0 1 0 1 0 1 0
          1 0 1 0 0 0 0 0 1 0 0 0 1 1 1 1 0 0 0 0 0 0 1 1 1 0 0 1 0 1 1 1 0 0 1 0 0
          1 0 0 0 0 0 1 1 1 0 1 0 0 0 1 1 0 1 0 1 0 0 1 0 0 0 0 0 0 0 1 0 1 0 1 1 0
          1 1 0]

         [1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 1 1 1 0 0 1 0 0 1 0 1 0 1 0 1 0 1 0
          1 0 1 1 0 1 0 0 1 0 0 0 1 1 1 1 0 0 0 0 0 0 1 1 1 0 0 1 0 1 1 1 0 0 1 0 1
          1 0 0 0 0 0 1 1 1 0 1 0 0 0 1 1 0 1 0 1 0 0 1 0 0 0 0 0 0 0 1 0 1 0 1 1 0
          1 1 0]
```

3.6. ALGORITHMS USED

3.6.1. Logistic Regression

It is a statistical model used to model a binary dependent variable. Mathematically, a binary model has two possible outcomes- 0 or 1. Logistic Regression uses an equation much like linear regression as the representation. A dichotomous variable is used to assess the outcome. To tie the input and output together, a “logit‟ function is used that represents a linear combination of variables as a map, as a result, a probability distribution with a domain of 0 to 1 is obtained.

$$\text{Log}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1$$

The following is the standard equation of logistic regression:

$$y = \frac{e^{b0+b1*x}}{(1 + e^{b0+b1*x})}$$

The predicted output is y, the intercept term is b0, and the input value coefficient is b1 (x).

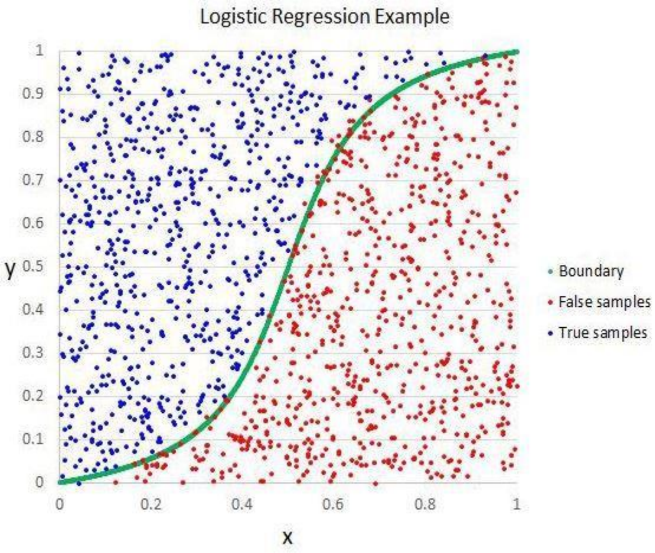


Figure 3.6.1- Example of Logistic Regression

3.6.2. Decision Tree

The Decision Tree is an algorithm that falls under supervised learning and is used in both casesregression and classification. They use tree representation to deal with the problem. Each leaf node represents a class label, whereas the inside node represents properties. CART stands for decision tree algorithms (Classification and Regression Trees).

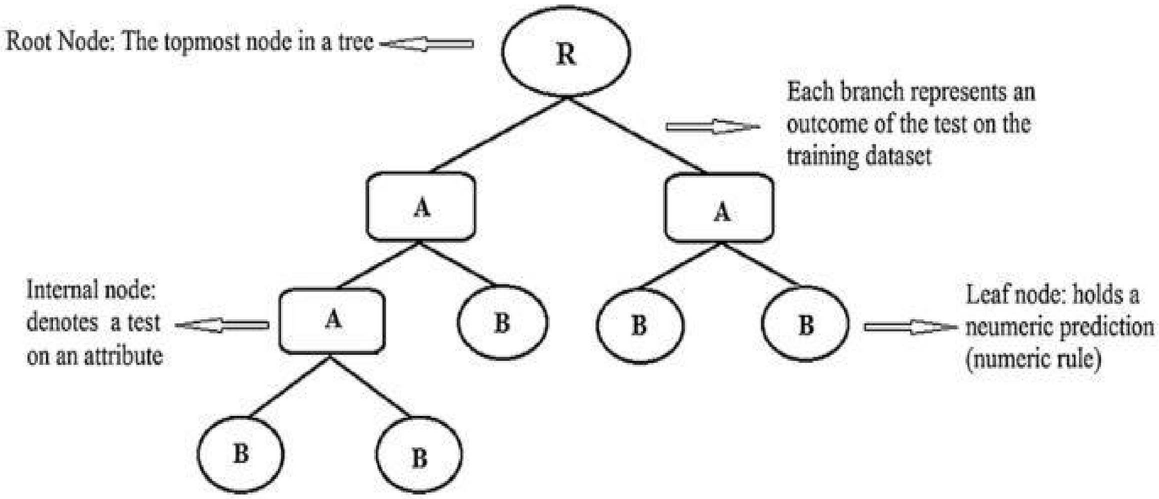

Root Node: The topmost node in a tree
R
Each branch represents an outcome of the test on the training dataset
A
A
Internal node: denotes a test on an attribute
A
B
B
B
Leaf node: holds a neumeric prediction (numeric rule)
B
B

Figure 3.6.2.- Decision Tree Classifier representation

The main objective is to find a suitable attribute Decision Tree works on the following 4 algorithms:
a. Gini Index
b. Chi-Square
c. Information Gain
d. Reduction in Variance

3.6.3. Random Forest Algorithm

Random Forest algorithm, like its name suggests is a collection of individual decision trees which operate as an Ensemble. An Ensemble means the final output depends on the outputs of the individual decision trees. Rather than just just calculating the average pf the output of trees, random forest implements two key methods:

i. Random Sampling of training data- In the training period, each tree learns a selection of data points at random. Forecasts are made by averaging the predictions of each decision tree at a test time. Bagging, short for bootstrap aggregating, is the process of training each individual learner on distinct bootstrapped subsets of the data and then averaging the predictions.

ii. When separating the nodes, random subsets of characteristics are used- This is one of the main concept in which, for each node, a subset of all features is taken into account.

CHAPTER 4

RESULT AND DISCUSSION

Random Forest, Decision Tree and Logistic Regression algorithm are implemented and has been proposed. Open source machine learning libraries like numpy, pandas and scikit learn are used in the project. Jupyter which is an open source web application was used to implement and run the program. The results presented depicts that, the recall performance metric of Random Forest is the best and Decision Tree has 100% classification accuracy in training data set but Random Forest Classifier is the most accurate, precise, and has the highest F1 score over Decision Tree and Logistic Regression. Finally Random Forest Classifier is at 97.36% accuracy and Logistic Regression is at 96.49% accuracy and Decision Tree is at 93.85% accuracy.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

The most frequent malignancy is breast cancer. A woman chosen at random has a twelve percent probability of being diagnosed with the condition. As a result, early identification of breast cancer can save many lives. The proposed model in this work is a comparative analysis of various machine learning methods for breast cancer detection. The performance of machine learning algorithms approaches was compared using the Wisconsin Diagnosis Breast Cancer data set. Each of the algorithms was shown to have an accuracy of more than 94 percent in determining whether a tumour was benign or malignant. It is found that Logistic Regression and Random Forest prove to be better in efficiency on the basis of accuracy, precision and F1 score compared to other algorithms. Thus, in cancer research, supervised machine learning approaches will be highly helpful in early diagnosis and prognosis of a cancer kind. This time, the confusion matrix reveals surprisingly good results; the neural network is misclassifying less in both classes, as evidenced by the values of the main diagonal and the accuracy value of around 97 percent. It signifies that a new, unseen case has a 97 percent chance of being correctly classified by the classifier.

Linear Discriminant Analysis, Factor Analysis, Isomap, and its variants are examples of dimensionality reduction approaches. The goal is to examine each one's benefits and drawbacks, as well as to compare and contrast their results individually and in combination.

REFERENCES

[1].Joseph A. Cruz, David S. Wishart, “Application of Machine Learning in Cancer Prediction and Prognosis”, Departments of Biological Science and Computing Science, University of Alberta Edmonton, AB, Canada T6g 2E8, 2006.

[2].Shubham Sharma, ArchitAggarwal, TanupriaChoudhury, “Breast Cancer Detection using Machine Learning Algorithms”, University Of Petroleum and Energy Studies, Dept of Informatics, School of Computer Science, Dehradun Amity University Uttar Pradesh, 2018 IEEE

[3].WenbinYue, Zidong Wang, Hongwei Chen, Annette Payne, Xiaohui Liu, “Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis”, Department of Computer Science, Brunel University London, Uxbridge Middlesex UB8 3 PH, UK, School of Mathematics, Southeast University , Nanjing 210096 China, 2018.

[4].KonstantinaKourou, Themis P. Exarchos, Konstantinos P. Exarchous, Michalis V. Karamouzis, Dimitrios I. Fotiadis,”Machine Learning Application in Cancer Prognosis and prediction”, Unit of Medical Technology and Intellegent Information Systems, Dept. of Material Science and Engineering, University of Ioannina, Greece, IMBB-FORTH Dept. of Biomedical Research, Greece, Molecular Oncology Unit, Department of Biological Chemistry, Medical school, University of Athens, Athens, Greece, 2015.

[5].Mishra, S., Sarkar, U., Taraphder, S., Datta, S., Swain, D., &Saikhom, R. et al. (2017). Multivariate Statistical Data Analysis- Principal Component Analysis (PCA). International Journal of Livestock Research, 7(5), 60-78. http://dx.doi.org/10.5455/ijlr.20170415115235

[6].S. Ouyang and Z. Bao, “Fast principal component extraction by a weighted information criterion,” IEEE Transactions on Signal Processing, vol. 50, no. 8, pp. 1994–2002, 2002.

[7].Tom Howley, Michael G. Madden, Marie-Louise O‟Connell and Alan G. Ryder "The Effect of Principal Component Analysis on Machine Learning Accuracy with High Dimensional SpectralData" National University of Ireland, Galway, Proceedings of AI-2005, 25th International Conference on Innovative Techniques and Applications of Artificial Intelligence, Cambridge, Dec 2005.

[8].O. L. Mangasarian, W. N. Street, and W. H. Wolberg, “Breast cancer diagnosis and prognosis via linear programming,” *Operations Research*, vol. 43, no. 4, pp. 570–577, 1995.

[9]. A. J. Cruz and D. S. Wishart, “Applications of machine learning in cancer prediction and prognosis,” *Cancer Informatics*, vol. 2, pp. 59–77, 2006.

[10]. S. Gc, R. Kasaudhan, T. K. Heo, and H.D. Choi, “Variability Measurement for Breast Cancer Classification of Mammographic Masses,” in *Proceedings of the 2015 Conference on research in adaptive and convergent systems (RACS)*, Prague, Czech Republic, 2015, pp.177–182.

[11]. C. Wang, W. Wang, S. Shin, and S. I. Jeon, “Comparative Study of Microwave Tomography Segmentation Techniques Based on GMM and KNN in Breast Cancer Detection,” in *Proceedings of the 2014 Conference on Research in Adaptive and Convergent Systems (RACS '14)*, Towson, Maryland, 2014, pp.303–308.

[12]. C. L. Chowdhary, and D. P. Acharjya, “Breast Cancer Detection using Intuitionistic Fuzzy Histogram Hyperbolization and Possibilitic Fuzzy cmean Clustering algorithms with texture feature based Classification on Mammography Images,” in *Proceedings of the International ConferenceonAdvancesinInformationCommunicationTechnology& Computing*, Bikaner, India, 2016, pp.1–6.

[13]. S. Aminikhanghahi, S. Shin, W. Wang, S. I. Jeon, S. H. Son, and C. Pack, “Study of wireless mammography image transmission impacts on robust cyber-aided diagnosis systems,” *Proc. 30th Annu. ACM Symp. Appl. Comput. - SAC ’15*, pp. 2252–2256,2015.

[14]. S.G.Durai,S.H.Ganesh,andA.J.Christy,“NovelLinearRegressive Classifier for the Diagnosis of Breast Cancer,” *In Computing and Communication Technologies (WCCCT), 2017 World Congress on* 2017.

[15]. H.Wang,andS.W.Yoon,“Breastcancerpredictionusingdatamining method,” *IIE Annu. Conf. Expo 2015*, pp. 818–828,2015.

[16]. S. Hafizah, S. Ahmad, R. Sallehuddin, and N. Azizah, “Cancer Detection Using Artificial Neural Network and Support Vector Machine:AComparativeStudy,”*J.Teknol*,vol.65,pp.73–81,2013.

[17]. A.T.Azar,andS.A.El-Said,“Performanceanalysisofsupportvector machines classifiers in breast cancer mammography recognition,” *Neural Comput. Appl.*, vol. 24, no. 5, pp. 1163–1177,2014.

[18]. C.Deng,andM.Perkowski,“ANovelWeightedHierarchicalAdaptive Voting Ensemble Machine Learning Method for Breast Cancer Detection,”*Proc.Int.Symp.Mult.Log.*,vol.2015–Septe,pp.115–120, 2015.

[19]. U.Rehman,N.Chouhan,andA.Khan,“DiverseandDiscriminative Features Based Breast Cancer Detection Using Digital Mammography,” *2015 13th Int. Conf. Front. Inf. Technol.*, pp. 234– 239,2