

**PERFORMANCE ANALYSIS OF CLASSIFICATION ALGORITHMS FOR
BREAST CANCER PREDICTION**

BY

MD. SHIHAB UDDIN
ID: 171-15-8836

MD. TAREQ HASAN
ID: 171-15-9213

DIPTO SARKAR
ID: 171-15-8626

This Report Presented in Partial Fulfillment of the Requirements for the
Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Mr. Narayan Ranjan Chakraborty
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Mr. Shah Md Tanvir Siddiquee
Assistant Professor
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

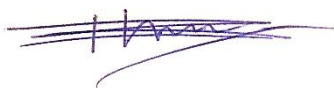
DHAKA, BANGLADESH

JUNE 2021

APPROVAL

This Project titled “**Performance Analysis of Classification Algorithms for Breast Cancer Prediction**”, submitted by Md. Shihab Uddin, Md. Tareq Hasan and Dipto Sarkar to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 3 June 2021.

BOARD OF EXAMINERS



Dr. Touhid Bhuiyan

Professor and Head

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman



Subhenur Latif

Assistant Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Md. Abbas Ali Khan

Senior Lecturer

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



External Examiner

Dr. Md Arshad Ali

Associate Professor

Department of Computer Science and Engineering
Hajee Mohammad Danesh Science and Technology
University

DECLARATION

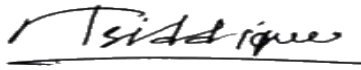
We hereby declare that, this project has been done by us under the supervision of **Mr. Narayan Ranjan Chakraborty, Assistant Professor, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:



Mr. Narayan Ranjan Chakraborty
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:



Mr. Shah Md Tanvir Siddiquee
Designation
Department of CSE
Daffodil International University

Submitted by:



Md. Shihab Uddin
ID: 171-15-8836
Department of CSE
Daffodil International University

Tareq

Md. Tareq Hasan

ID: 171-15-9213

Department of CSE

Daffodil International University

Dipto

Dipto Sarkar

ID: 171-15-8626

Department of CSE

Daffodil International University

ACKNOWLEDGEMENT

At first, we would like to express our heartiest thanks and gratefulness to almighty Allah for his divine blessing makes us possible to complete the final year thesis successfully.

We really grateful and wish our profound our indebtedness to **Mr. Narayan Ranjan Chakraborty**, Assistant Professor, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Machine Learning*” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Dr. Touhid Bhuiyan**, Professor, and Head, Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

In modern time, we can say that cancer is one of the threatening diseases ever. We can see from the world figures that there are many major health concerns in the current situation. The mortality rate of cancer patients is very high as we research cases of multiple cancers. Patients have some signs that are characteristic of early-stage cancer. The breast cancer survival rate can be improved if the breast cancer tumor can be detected in the early stage and all control functions can be reduced so that it doesn't expand or migrate to other organs. The survival rate can increase. And if we see multiple stages of cancer, we will see that breast cancer is among the most prominent in the current scenario. If we research right, we can see those multiple types of breast cancer data sources exist.

TABLE OF CONTENTS

CONTENTS	PAGE
Acknowledgement	vi
Abstract	vii
List of Figures	xi
List of Tables	xii
CHAPTER 1: Introduction	1-5
1.1 Introduction	1
1.2 Motivation	2
1.3 Rational of the Study	3
1.4 Research Questions	4
1.5 Expected Output	5
1.6 Report Layout	5
CHAPTER 2: Background	6-10
2.1 Terminologies	6-7
2.2 Related Work	8
2.3 Research Summary	9
2.4 Challenges	10

CHAPTER 3: Methodology	11-18
3.1 Introduction	11-12
3.2 Research Subject and Instrumentation	13
3.3 Data Collection and Data preprocessing	14
3.3.a Get Dataset	15
3.3.b Check and Remove Unnecessary Columns	15
3.3.c Removing Null Values and Replacing Values	15
3.3.d Purified Data	16-17
3.4 Statistical Analysis	18-19
CHAPTER 4: Implementation	20-34
4.1 Support Vector Machine (SVM)	20-22
4.2 Random Forest Classifier	23-24
4.3 K Nearest Neighbor	25-28
4.4 Gaussian Naïve Bayes	29-30
4.5 Logistic Regression	31
4.6 Decision Tree	32-34

CHAPTER 5: Experimental Results, Discussion and Conclusion	
5.1 Introduction	35-45
5.2 Experimental Result	35
5.3 Conclusion	36-44
	45
APPENDICES	46
REFERENCES	47-49

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1.1: Workflow for breast cancer detection	12
Figure 3.3.1: Dataset pre-process steps	14
Figure 3.3.2: Before pre-processing data	16
Figure 3.3.3: After pre-processing data	17
Figure 3.4.1: Dataset preview	18
Figure 4.1.1: SVM figure	20
Figure 4.1.2: How SVM works	21
Figure 4.1.3 Managing nonlinear and indistinguishable planes	22
Figure 4.2.1 Steps of how Random Forest classifier works	24
Figure 4.3.1 KNN flowchart	26
Figure 4.3.2 The process of KNN algorithm	27
Figure 4.3.3 The process of KNN algorithm	28
Figure 4.4.1 Naïve Bayes working process	29
Figure 4.5.1 Logistics Regression	31
Figure 4.6.1 Decision Tree	32
Figure 4.6.2 Working process of Decision Tree	33
Figure 5.2.1 SVM result	36
Figure 5.2.2 Decision Tree result	37
Figure 5.2.3 Gaussian Naïve Bayes result	38
Figure 5.2.4 KNN result	39
Figure 5.2.5 Random Forest result	40
Figure 5.2.6 Logistic Regression result	41
Figure 5.2.7 Performance comparison	43
Figure 5.2.8 Number of Benign and Malignant cases from the dataset	43
Figure 5.2.9 Breast Cancer Attributes correlation heatmap	44
Figure 5.2.10 General Gaussian Distribution	45

LIST OF TABLES

TABLES	PAGE NO
Table 1: Results of Different Algorithms	42

CHAPTER 1

Introduction

1.1 Introduction

Breast cancer is known is one of the dangerous and most common nowadays for women all across the world. If we check the global statistics, we can see that in the present situation there are a lot of significant issues according to health. If we study the cases of different cancers the death rate of cancer patients is quite high. In early-stage cancer, patients have some symptoms which are descriptive. The survival rate of breast cancer can be increased if the tumor of breast cancer can be identified in the early stage and if the tumor can be minimized all monitor properly so that it will not increase the size or spread through other organs. And if we see the evolution of different stages of different cancers, we can see breast cancer is one of the most common for the present situation. If we study properly, we can see that there are different kinds of data sources for breast cancer. Many researchers use different kinds of methods or techniques to identify and classify breast cancer. There is a lot of research work according to this topic, but most of the researchers used at most two or three algorithms in the research works. We are trying to increase the number of the algorithm in our research work and merge the results of different algorithms of machine learning. We will use here the machine learning technique to find out or identify breast cancer. If the best cancer patient can be identified through diagnosis, we can minimize the effect and it is sometimes possible to cure the breast cancer. We have known about different and rhythms of techniques to find and classify the breast cancer problem some of the most commonly used algorithms are support vector machines, decision tree, k-nearest neighbor algorithm, and so on. Most of the researchers used support vector machines and decision tree algorithm to identify breast cancer or predict if a patient has breast cancer or not.

1.2 Motivation

The plurality of new cancer cases and cancer-related deaths in women worldwide, according to global estimates, reflect Breast cancer is one of the most prevalent cancers in the world, making it a critical public health issue in society today. This article provides an analysis of the evolution in the health system of broad data and applies four research algorithms to a sample of breast cancer data. The goal of this investigation is to forecast breast cancer, the second leading death cause in the world among women, and to reduce the risk of death by means of a variety of machine-learning techniques early detection and prevention. The entitlement of humanity is a safe and disease-free life. Cancer is human beings' most extreme illness which can lead to death. So, to eradicate these pathogens and therefore to protect people from them, a conclusive approach must be sought. Breast cancer is known as one of the risky forms of cancer, especially for women. Early exams should take place regularly and a more responsive and successful diagnosis should be made to safeguard the lives of women. Gender, thick breast tissue, alcoholism, body weight, and radiation pollution are the primary causes of breast cancer. The probability of survival improves if this is detected at an early stage. Original breast cancer may be detected through machine learning and profound learning.

Breast cancer has always been the second most prevalent disease with the largest number of people living with breast cancer in both the developing and developed nations. Breast Cancer is distinguished by gene mutation, chronic pain, change in height, color and the appearance of the skin of the breast. Classification of breast cancer contributes pathologists to a standardized and definitive pronouncement, typically the most common classification is linear (mild cancer/malignant cancer). In the classification problem of breast cancer, machine learning strategies are still widely utilized. They are graded and are diagnosed with a high degree of consistencies.

1. Choosing the best breast cancer prediction classification
2. Comparison of various breast cancer dataset algorithms.
3. Identifying the most effective disease prediction algorithm.

Breast cancer is caused by a specific cell type or defect that can be shut down by the system that triggers a sloppy cell proliferation. Breast cancer, If the issue is not corrected after several months, cells with incorrect orders shape masses.

1.3 Rational of the study

We will represent the breast cancer and the classification throughout this segment initially.

Based on the class of the disease, Breast Cancer Classification seeks to assess which therapy can be effective or less aggressive. To ensure a positive prognosis, the classification of breast cancer includes nine features, which include:

1. Determines the layered structures;
2. Assess sample size and accuracy (Cell Size Uniformity);
3. Estimate cell equality and define median fluctuations because cancer cells appear to be different in their structure (Cell shape uniformity);
4. Cells of cancer disperse across the organ and normal cells are interconnected (Marginal Adhesion).
5. A uniformity measurement and stretched epithelial cells are a mark of malignancy.
6. Cytoplasm (Bare Nuclei) is not found in malignant lesions;
7. Describes the arrangement of the nucleus, which is uniform in benign cells. chromatin in tumors is frequently coarser (Bland Chromatin);
8. The nucleolus often becomes noticeable and very fragile in normal cells. There are more than few nucleoli (Normal Nucleoli) in cancer cells and it becomes even more predominant;
9. A measurement of the amount of mitosis emerging. The greater the benefit, the more malignancy is conceivable (Mitoses).

1.4 Research Questions

- What is Breast Cancer analysis?
- How Breast Cancer analysis works?
- What are the benefits of Breast Cancer analysis?
- What are the problems of Breast Cancer analysis using machine learning?
- How is to train and preprocess the Breast Cancer dataset?
- What are scopes in future research of this projects?
- How does the Model of different algorithms work?
- What is the machine learning techniques used to complete the research work?
- What are the challenges raised while doing this work?
- What is the way to get all the data?
- What parameters are used in the algorithms?
- How many algorithms did you used in this work?
- Which algorithm shows the best result and why?
- What are the major differences between this research work and others?
- What is the expected output form this research work?

1.5 Expected Output:

Since this is research and our goal is to go for Publication after getting a good accuracy, so we are planning to publish research paper in this related field. There are many researchers who work on this topic they have very good datasets to find out or credit if there is any sign of breast cancer or not. Some of the mainly e used random forest, K nearest neighbors' algorithm and much more different machine learning algorithms to find out if it can properly predict or detect breast cancer. Our main goal is to use all the different algorithms and merge it in in one research work to find out which machine learning techniques actually working more accurately than others. And which will be the most efficient way to find out breast cancer.

So, we can say that the main differences between the other researchers and ours are we are trying to find out the most accurate machine learning technique for detecting or finding if there is any chance of having breast cancer to any patient. So, there are to main things which will help us to find out breast cancer are: Malignant and Benign. These two things are important for detecting breast cancer. In this research work, we have the plan to implement 6 algorithms to compare and find the best algorithm for best machine learning technique which will help us to find out or detect breast cancer. We are expecting to get better accuracy. The algorithms which will give us the highest accuracy for near to the highest accuracy we will use does machine learning techniques for this breast cancer detection. If we get better accuracy, we have the plan to publish this research work to a journal or conference.

1.6 Report Layout:

In this report in the first chapter, we have discussed an overall view of the whole research work. In the second chapter, we have discussed the background studies which are related to this works. In the next chapter, we have discussed the research methodology. In the fourth chapter, we discussed which algorithms we used. In the last chapter, we have discussed the experimental results and conclusion.

CHAPTER 2

Background Studies

2.1 Terminologies:

In women worldwide, breast cancer is the most prevalent cancer. It has been researched thoroughly over the years. Breast cancer treatment has actually made it possible to make advances in other forms of cancer research. Since the first cancer discovery, how we treat breast cancer has improved in many respects. Yet for years other studies and therapies have been the same. For a long time, people have heard about breast cancer. The Edwin Smith Surgical Papyrus, for example, details breast cancer cases. In the ancient Greece, people made votive offers in the form of breast for the god of medicine. This medicinal text was written in 3,000-2,500 A.D the first century, in the first century, surgeons dealt using surgical incisions to remove tumors. Hippocrates explains the stage of breast cancer in the early 400s A.D. Therapy for breast cancer is increasingly specialized as doctors learn more about it. It is now regarded as a disorder of various models and ways of behaving on the body by subtypes. Specific genes are isolated and breast cancer identified as the beginning of tweak therapy alternatives. Even physicians can discover more about breast cancer through special examinations. For instance, a tumor component may be investigated in Oncotype DX gene feature analysis to ensure which genes are functioning there. Physicians will evaluate which patients can be treated with early breast cancer separately and who will require chemotherapy additionally.

They suspected that the termination of menstruation was associated to breast cancer. This interpretation may have led to cancer being linked with later adults. Medical development was related to new religious philosophies in the early Middle Ages. Christians concluded the procedure was barbaric and favored a remedy in confidence. In the meantime, Islamic physicians studied Greek medical texts to find out a little more about breast cancer. In the Renaissance the surgery re-examined the human body. The Scottish founder of investigational surgery, John Hunter, described lymph as a cause of

breast cancer. Lymph is the white blood cells that hold fluid throughout the body. The surgeons still carrying out lumpectomies, but no anesthesia was still required. Surgeons wanted to make a difference quickly and reliably.

Any of the identified breast cancer risk factors are as follows. Most forms of breast cancer, nevertheless, cannot be associated with any specific factor. Evaluate the particular risk with a doctor. Age, Levels, Women get older are more likely to have breast cancer. In women over the age of 50 about 80% of breast cancers exist. Breast cancer family medical history. A woman who is at elevated possibility of acquiring breast cancer in her other breast has one breast. Breast cancer chronic illness. If your mother, sister, or daughter has breast cancer, especially when she is young, a woman is more at risk of breast cancer (before 40). It can also affect the possibility of other members with breast cancer. Factors biological. The risk of developing breast cancer over your lifetime is greater for women of some genetic mutations, particularly alterations in BRCA1 and BRCA2 genes. Other variations of the gene can also increase the risk of breast cancer. The background of childhood and menstruation. The older a woman, the greater her chance of breast cancer is when she has her first child. Women who first menstruate earlier (until 12 years) women who go into menopause later (after 55 years of age) women who have never had children are also at heightened disadvantage.

2.2 Related Work:

Breast Cancer is one of the threatening diseases in the world after lungs and blood cancer. Many women are facing this terrible disease it may cause severe damage to the body. Many researchers are working on this topic using various techniques and algorithms. Some researchers used machine learning technique to classify breast cancer. They used two separate classifiers: Naive Bayes (NB) classifier and k nearest neighbor (KNN) for this. They have very good results; NB algorithm gives 96.19% and KNN gives 97.51% accuracy. [3] From another research paper published in 2016 we have learnt that some researchers used Support Vector Machine (SVM), KNN, Decision Tree (C4.5), and NB algorithms to find out the risks for breast cancer with these algorithms. SVM algorithm gives 97.13% accuracy to predict the risk of breast cancer. [2] In another research work the researchers used an automatic MRI dataset to identify if there is any breast tumor or not. They used feature extraction discrete wavelet transform (DWT) from the images of the dataset. After feature extraction they used SVM algorithm to detect breast tumor. The model gives 98.03% accuracy. [12] Another group of researchers used Wisconsin database to do breast cancer classification to get more efficient approaches. They use SVM, KNN, CNN classifiers, Logistic Regression, Random forest algorithms and the accuracy they obtain is up to 98% to 99%. [7] In another research work in 2018 some researchers used data mining techniques to create a survey in breast cancer. They used clustering algorithms to make the survey.[8] In August 2020 a group of researchers review different types of machine learning techniques to do prediction on breast cancer. They used NB=98.2%, J48=98.5% and KNN=98.8% accuracies respectively. [9] On an Indian conference in 2020 some researchers used XG Boost and Random forest algorithms to predict the risk of breast cancer. They used 275 instances with 12 different features to complete the analysis. They got 74.73% accuracy using the Random Forest algorithm and using XG Boost they got 73.63% accuracy. [10] On August 2020, some researchers compared SVM algorithm with five others ML algorithms, they are: KNN, LDA (Linear Discriminant Analysis), NB, LR, Decision tree classifier (CART) to evaluate soft classification techniques for breast cancer detection. But they got the

highest accuracy by using SVM algorithm. [15] In 2018 a research work was published in a conference in Morocco, in that research work the researcher used SVM algorithm to find breast cancer using machine learning and got 97.9% accuracy.[29]

2.3 Research Summary:

Our research work we have the Goal to use a data set to which is related to breast cancer and after having the data set you will find out what part of the data set will be needed on different machine learning algorithm. According to different machine learning algorithm there are different models and techniques for each of them. And we will try 6 different machine learning algorithms to find out the how much accurately these algorithms can detect breast cancer. Our main goal is to find out if any patient has the factors which are related to breast cancer. Then we will also evaluate how much accurately the machine learning algorithms works for the data set. Then we will compare each and every algorithm and find out which machine learning algorithm will give us the best prediction to detect breast cancer. There are many researchers who was with some particular algorithms but we will try to combine 6 different algorithms in one research work which will help us to compare and understand each algorithm and how it works. And what factors will change the result of predicting breast cancer. Different researchers have done their research work on the basis of different types of data. Some researchers work on different symptoms and train the data set to match how many symptoms matched with the data set and the patient then the machine gives the prediction that if the patient have breast cancer or not.

2.4 Challenges

There are some challenges we have to face while doing this research work. One of the main challenges was to find a suitable dataset as per our demand. Then we have to improve the dataset. our main goal is to find which datasets are used to detect or predict if any patient has breast cancer or not. There are many data set which are used in different research work but not all of them are used to predict or detect breast cancer. So, we have to find the usable data set and modify it. We have collected the data from Kaggle and different researchers. Then we have to process the data and change some value according to our research work because we have to prepare the data set for multiple machines learning algorithms. As we already mentioned that we are trying to use six different machine learning algorithms in one research work that means we have to prepare the data set for all those six machines learning algorithms. After having the data set, we have to change some values and remove some columns and values. After that we have to identify which 6 different algorithms we will use and how it works with other researchers. One of the most popular data set is Wisconsin breast cancer dataset.[9][5][4] Here we added some reference those researchers work with Wisconsin breast cancer dataset. A good amount of data is used for this research work.

CHAPTER 3

Research Methodology

3.1 Introduction

In this segment, we will address the whole technique of the research work. Each analysis has a particular method of solving. At first, we have collected the data set from Kaggle Wisconsin breast cancer dataset. Then we do the dataset and process the data after that we have to remove some null values and some columns which we do not need here for our research work. Then we select which machine learning algorithm will be used. We have already mentioned that we are using six different machine learning algorithms so as per the algorithm and the model be built, we have to prepare a data set to fit in the model and train our data set. Feature selection is working on that basis. After that we have partition the data and split it in training set and testing set. Which is also known as train data set and test data set. After that we fit it in different machine learning algorithm models and train the data with the train data set, after that we have portion of data for testing our model. And evaluate the model and get the accuracy. With this accuracy we can say that if anyone have any chance to get attack by breast cancer. We have followed our common workflow flowchart that will give us an overview but we will discuss some of the algorithms broadly with their equations and with some diagrams to understand how the algorithm works.

The workflow of the entire research work is given below, which provides a brief summary of the overall research work.

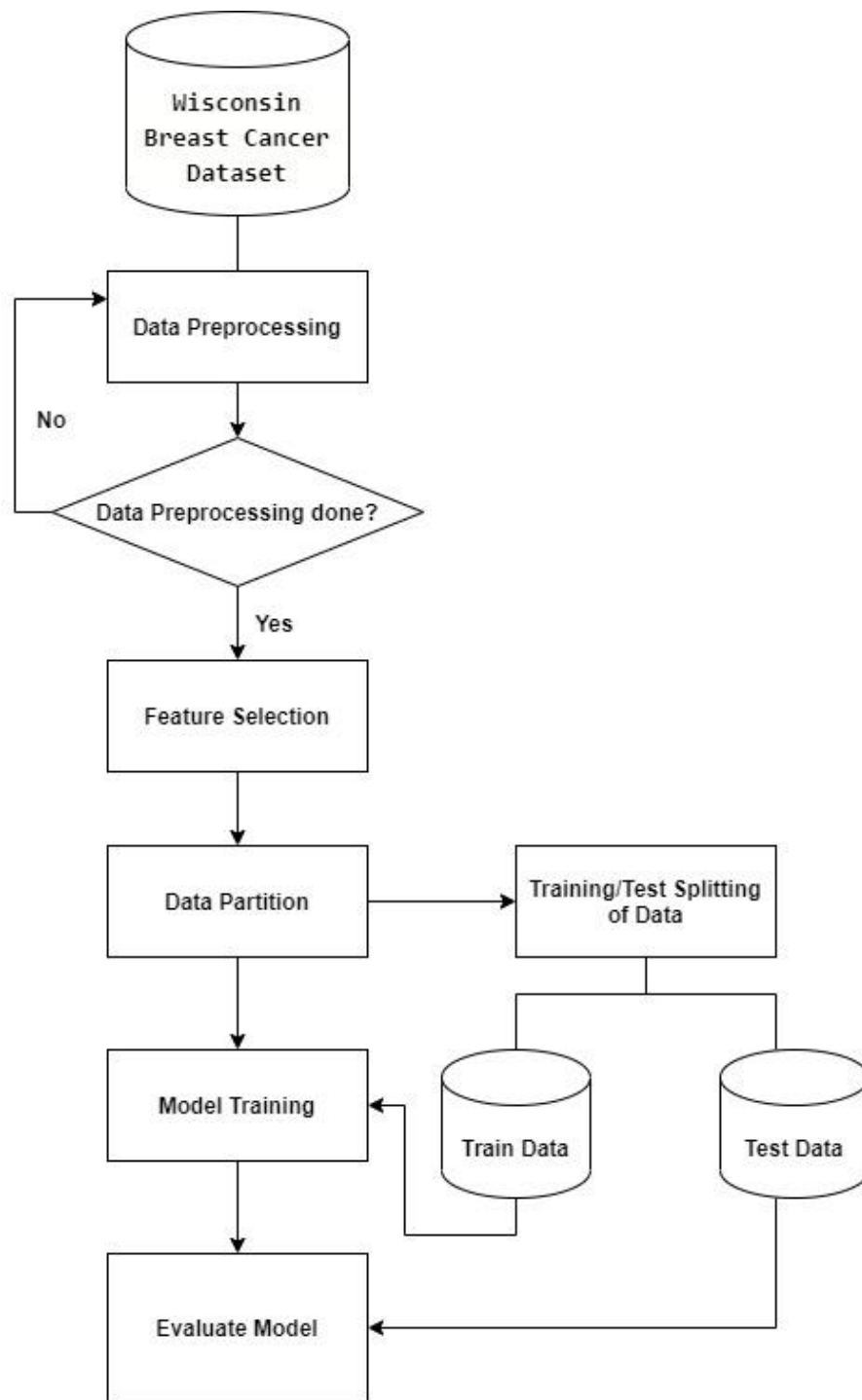


Figure 3.1.1 Workflow for Breast cancer detection

3.2 Research Subject and Instrumentation

In the first place, we have discussed the process of detecting breast cancer with the conceptual and theoretical process. Machine learning model needs high configuration pc with GPU and others instrument. Now a list is given below of the required instrument for this model.

Hardware and Software:

- Intel Core i5 8th generation
- 1 TB Hard Disk Drive
- 8GB RAM

Development Tools:

- Windows 10
- Python 3.8
- Pandas
- Seaborn
- Matplotlib
- NumPy
- Scikit-Learn

3.3 Data collection and Data preprocessing

We have collected Dataset from Kaggle and as per our need we have changed the data set a little bit. After that for each algorithm we have to use the dataset. As we already told that we are using 6 algorithms, so each algorithm works differently, we have to preprocess the data properly to fit the data set in our model.

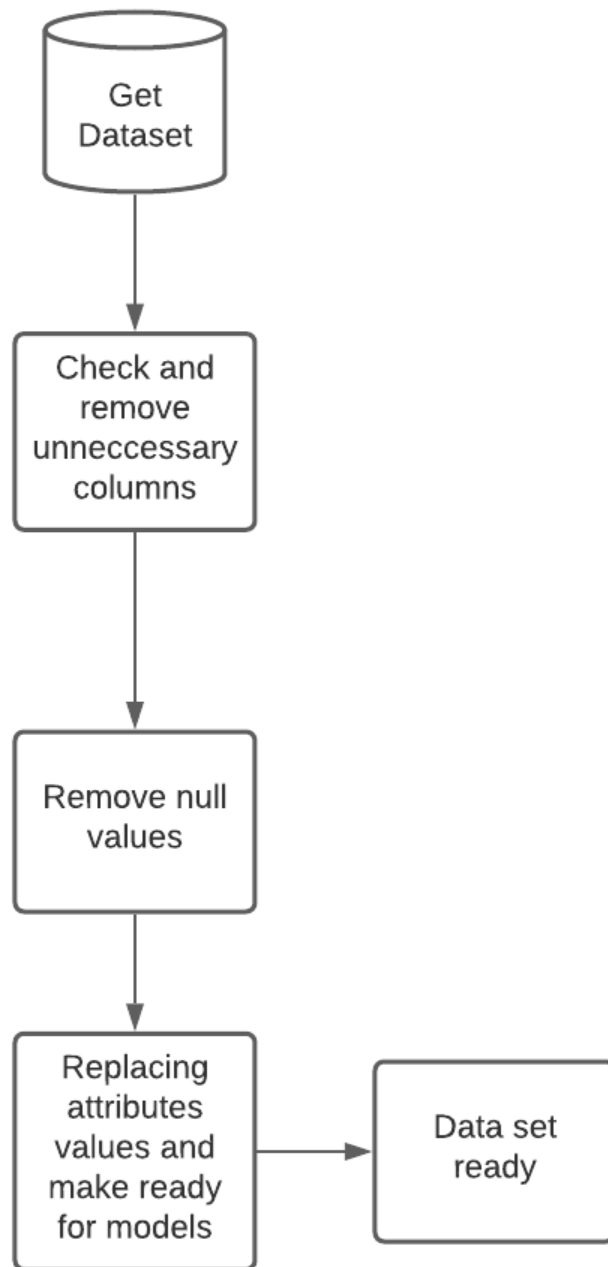


Figure 3.3.1: Dataset pre-process steps

3.3.a Get Dataset:

As previously we have told that we have collected data from different sources but different researchers work with different kind of data sets. So, we have to pick the data set carefully because we want to use one data set for several algorithms to detect or predict if any patient all human being has breast cancer or not. That is the reason we have to choose the dataset carefully and we have to preprocess the data set properly.

3.3.b Check and remove unnecessary columns:

The removal of unnecessary columns is a very important thing for data preprocessing. We have to check if there is any null value or not. The null value actually decreases the accuracy level, so removing the null value will give us a better result because if there is no null value it will help to train the data set properly and give us more accurate result. And we have to remove some unnecessary columns, the reason is we do not need all the columns to fit in our model.

3.3.c Removing null values and Replacing values:

As we have already told that removing null value from data set is important because if we do not remove the null values when we use different algorithms while Training it will not properly trained because of these null values. And for these null values sometimes the model of the algorithm will generate less accurate result. Sometimes some data may miss, or sometimes the data may not enter properly so that kind of things should be removed. If we remove all the null values properly then attack training when we test the algorithm with this dataset, we can see the higher accuracy. In our data set there is an attribute for column name diagnosis there are two kind of levels there one is 'M':1,'B':0. So we have to replace the M values with 1 and B values with 0.

3.3.d Purified data:

After following all the steps of data preprocessing like Check and remove unnecessary columns, removing null values, finally the data we get is pea processed or purified data. These preprocess data will look clean than the original data. The original data may contain different kind of things but after pre-processing it gets a clean look which is easier to be trained for machines. So, the preprocess data is actually what we are giving to the machine to be trained.

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal_dimension_mean	radius_se
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419	0.07871	1.0950
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069	0.05999	0.7456
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597	0.09744	0.4956
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809	0.05883	0.7572

compactness_worst	concavity_worst	concave points_worst	symmetry_worst	fractal_dimension_worst	Unnamed: 32
0.6656	0.7119	0.2654	0.4601	0.11890	NaN
0.1866	0.2416	0.1860	0.2750	0.08902	NaN
0.4245	0.4504	0.2430	0.3613	0.08758	NaN
0.8663	0.6869	0.2575	0.6638	0.17300	NaN
0.2050	0.4000	0.1625	0.2364	0.07678	NaN

Fig 3.3.2: Before Pre-processing data

```

1 data = data_raw.copy()
2 data.drop(['id', 'Unnamed: 32'], axis=1,inplace=True)

1 data.head()

```

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concav
0	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	
1	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	
2	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	
3	M	11.42	20.38	77.58	386.1	0.14250	0.28390	
4	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	

```

1 data['diagnosis'] = data['diagnosis'].map({'M':1, 'B':0})
2 data.head()

```

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concav
0	1	17.99	10.38	122.80	1001.0	0.11840	0.27760	
1	1	20.57	17.77	132.90	1326.0	0.08474	0.07864	
2	1	19.69	21.25	130.00	1203.0	0.10960	0.15990	
3	1	11.42	20.38	77.58	386.1	0.14250	0.28390	
4	1	20.29	14.34	135.10	1297.0	0.10030	0.13280	

Fig 3.3.3: After Pre-processing data

3.4 Statistical Analysis:

1. We use a big amount of data. These data have 31 subsections such as ['diagnosis', 'radius_mean', 'texture_mean', 'perimeter_mean', 'area_mean', 'smoothness_mean', 'compactness_mean', 'concavity_mean', 'concave points_mean', 'symmetry_mean', 'fractal_dimension_mean', 'radius_se', 'texture_se', 'perimeter_se', 'area_se', 'smoothness_se', 'compactness_se', 'concavity_se', 'concave points_se', 'symmetry_se', 'fractal_dimension_se', 'radius_worst', 'texture_worst', 'perimeter_worst', 'area_worst', 'smoothness_worst', 'compactness_worst', 'concavity_worst', 'concave points_worst', 'symmetry_worst', 'fractal_dimension_worst']. A short view of our dataset is given below.

```
"id","diagnosis","radius_mean","texture_mean","perimeter_mean","area_mean","smoothness_mean","compactness_mean","concavity_mean","concave
points_mean","symmetry_mean","fractal_dimension_mean","radius_se","texture_se","perimeter_se","area_se","smoothness_se","compactness_se","concavity_se","concave
points_se","symmetry_se","fractal_dimension_se","radius_worst","texture_worst","perimeter_worst","area_worst","smoothness_worst","compactness_worst","concavity_worst","concave points_worst","symmetry_worst","fractal_dimension_worst",
8421302,"17.59,18.38,122.8,1.001,0.1104,0.2776,0.3001,0.1471,0.2419,0.07071,1.085,0.8053,0.589,153.4,0.006399,0.04904,0.05373,0.01507,0.03003,0.006193,25.38,17.33,184.6,2019,0.1622,0.6056,0.7119,0.2654,0.4601,0.1189
842517,"1,20.57,17.77,132.9,1.1326,0.00474,0.07864,0.0869,0.07017,0.1812,0.05607,0.5455,0.7339,3.398,74.08,0.005225,0.01308,0.0186,0.0134,0.01389,0.003532,24.99,23.41,158.8,1956,0.1238,0.1866,0.2416,0.186,0.275,0.08902
8430093,"11.49,21.25,138,1.086,0.1599,0.1974,0.1279,0.2069,0.05999,0.7456,0.7869,4.585,94.83,0.00615,0.04006,0.03832,0.02058,0.0225,0.004571,23.57,25.53,151.5,1708,0.1444,0.4045,0.4894,0.243,0.3613,0.08758
8434301,"11,41.20,38.77,58.386,1,0.1425,0.2839,0.2414,0.1052,0.2597,0.09744,0.4056,1.156,3.445,27.13,0.00911,0.07458,0.05661,0.01867,0.05963,0.009208,14.01,26.5,98.07,567.7,0.2080,0.0663,0.6869,0.2575,0.6638,0.173
84358402,"18,20.29,14.34,135.1,1297,0.1009,0.1328,0.198,0.1043,0.1089,0.05083,0.7572,0.7013,5.438,94.44,0.01149,0.02461,0.05688,0.01805,0.01756,0.005115,22.54,16.67,152.1,1575,0.1374,0.205,0.4,0.1625,0.2364,0.07678
843786,"12.45,15.7,82.57,477.1,0.1278,0.08089,0.2087,0.07613,0.2345,0.0962,2.217,27.19,0.00751,0.03345,0.03672,0.01137,0.02165,0.005082,15.47,23.75,103.4,741.6,0.1791,0.5249,0.5355,0.1741,0.3985,0.1244
844359,"18,25.19,98,119.6,1040,0.09463,0.189,0.1127,0.074,0.1794,0.05742,0.4467,0.7732,3.18,53.91,0.004314,0.01382,0.02254,0.01039,0.01369,0.002179,22.88,27.66,153.2,1608,0.1442,0.2576,0.3704,0.1932,0.3063,0.08388
84458202,"13,71.20,83,90.2,577.9,0.1189,0.1645,0.09366,0.05985,0.2196,0.07451,0.5835,1.377,3.856,58.96,0.008085,0.03029,0.02488,0.02448,0.01486,0.005412,17.06,28.14,110.6,897,0.1654,0.3682,0.2678,0.1556,0.3196,0.1151
844981,"13,21.82,87.5,519.8,0.1273,0.1932,0.1859,0.09353,0.235,0.07389,0.3063,1.002,2.406,24.32,0.005731,0.03582,0.03553,0.01226,0.02143,0.003749,15.49,30.73,106.2,739.3,0.1703,0.5401,0.539,0.206,0.4378,0.1072
84501001,"12,46.24,84.97,475.9,0.1186,0.2396,0.2273,0.08543,0.203,0.08243,0.2976,1.599,2.039,23.94,0.007149,0.07217,0.07743,0.01432,0.01789,0.01008,15.09,40.68,97.65,711.4,0.1853,1.058,1.105,0.221,0.4366,0.2075
845636,"16,82,23.24,102.7,797.0,0.08206,0.06669,0.03299,0.03323,0.1528,0.05697,0.3795,1.187,2.466,40.51,0.004029,0.009269,0.009269,0.01101,0.007591,0.0146,0.003042,19.19,33.88,123.8,1150,0.1181,0.1551,0.1459,0.09975,0.2948,0.08452
84610002,"15,78.17,24.8,132.4,0.0871,0.1292,0.09954,0.06066,0.1842,0.06802,0.5058,0.5049,3.564,54.16,0.005771,0.04061,0.02791,0.01282,0.02008,0.004144,20.42,27.28,136.5,1299,0.1396,0.5689,0.3905,0.181,0.3792,0.1048
846226,"19,17,24.8,132.4,0.1123,0.0974,0.2458,0.2065,0.1118,0.2397,0.078,0.9535,3.568,11.07,116.2,0.003139,0.008297,0.00889,0.0409,0.04484,0.01204,20.36,29.54,151.7,1332,0.1837,0.3803,0.3639,0.1767,0.3176,0.1023
846381,"15,85,23.95,103.7,782.7,0.08401,0.1002,0.09338,0.05364,0.1847,0.05338,0.4033,1.078,2.983,36.58,0.009769,0.03126,0.05051,0.01992,0.02981,0.003802,16.84,27.66,112.676,5,0.1131,0.1924,0.2322,0.1119,0.2889,0.06387
8467401,"13,73.12,41.63,96.7,578.3,0.1131,0.2291,0.2128,0.08925,0.2069,0.07682,0.2121,1.169,2.061,19.21,0.006429,0.05936,0.05981,0.01638,0.01961,0.008993,15.83,32.01,108.6,697.7,0.1651,0.7725,0.6943,0.2208,0.3596,0.1431
84799002,"14,54.27,54.96,73.658,0,0.1139,0.1595,0.1639,0.07364,0.1303,0.07077,0.37,1.033,2.879,31.55,0.005687,0.0434,0.04741,0.0109,0.01857,0.005466,17.46,37.13,124.1,943.2,0.1678,0.6577,0.7026,0.1712,0.4210,0.1341
848406,"14,68,20,13,94.74,684.5,0.09067,0.072,0.07395,0.05359,0.1586,0.05922,0.4717,1.34,3.195,45.4,0.005718,0.01162,0.02898,0.01109,0.0141,0.003085,19.07,30.88,123.4,1136,0.1444,0.1871,0.2914,0.1609,0.3029,0.00216
84862801,"16,13,20.68,180.1,798.8,0.117,0.2022,0.1722,0.1028,0.2164,0.07356,0.5862,1.079,3.854,54.18,0.007026,0.02501,0.03188,0.01297,0.01089,0.004142,20.96,31.48,136.8,1315,0.1709,0.4033,0.4784,0.2073,0.3706,0.1142
849014,"19,81,22.15,136,1260,0.09031,0.1027,0.1479,0.09498,0.1502,0.05395,0.7502,1.017,5.065,112.4,0.006494,0.01893,0.03391,0.01521,0.01356,0.001997,27.32,30.88,186.6,2398,0.1512,0.315,0.5372,0.2388,0.2768,0.07615
8510426,"8,13,54.34,36.87,46,566.3,0.09779,0.08129,0.06664,0.04781,0.1885,0.05766,0.2699,0.7086,2.058,23.56,0.008462,0.0146,0.02387,0.01315,0.0198,0.00215,11.19,26.99,7,711.2,0.144,0.1773,0.2398,0.1288,0.2977,0.07259
8510653,"8,13,88,15.71,85.63,520,0.1075,0.127,0.04568,0.0311,0.1967,0.06811,0.1852,0.7477,1.383,14.67,0.004997,0.01898,0.01698,0.00649,0.01678,0.002425,14.5,20.49,96.09,630.5,0.1312,0.2776,0.189,0.07283,0.3184,0.08183
8510824,"9,5,94,12,44,68,34,273.9,0.1024,0.06482,0.02956,0.02076,0.1815,0.06985,0.2773,0.5768,1.909,15.7,0.009686,0.01432,0.01985,0.01421,0.02027,0.002968,10.23,15.66,65.13,314.9,0.1324,0.1148,0.08867,0.06227,0.245,0.07773
8511113,"15,34,26,102.5,704.4,0.1073,0.2136,0.2077,0.09756,0.2521,0.07032,0.4388,0.7096,3.384,44.91,0.006789,0.05328,0.06446,0.02252,0.03672,0.004394,18.07,19.08,125.1,980.9,0.139,0.5954,0.6305,0.2393,0.4667,0.09946
851508,"21,16,23,04,137.2,1,044,0.08428,0.1022,0.1097,0.08632,0.1769,0.05278,0.6917,1.127,4.303,93.99,0.004728,0.01259,0.01715,0.01038,0.01083,0.001987,29.17,35.59,188,2615,0.1401,0.26,0.3155,0.2009,0.2822,0.07526
852552,"16,65,21,38,110,904.6,0.1121,0.1457,0.1525,0.0917,0.1995,0.0633,0.1868,0.9017,5.455,102.6,0.008048,0.01882,0.02741,0.0113,0.01468,0.002081,26.46,31.56,177,2215,0.1805,0.3578,0.4095,0.2093,0.3613,0.09564
852633,"17,14,16,4,116,912.7,0.1186,0.2276,0.2229,0.1401,0.304,0.07433,1.046,0.976,7.276,111.4,0.008029,0.03799,0.03732,0.02397,0.02308,0.007444,22.25,21.4,152.4,1461,0.1545,0.3949,0.3853,0.255,0.4866,0.1059
852783,"14,58,21,53,97.41,644.8,0.1054,0.1868,0.1425,0.08783,0.2252,0.06924,0.2545,0.9832,1.11,21.05,0.004452,0.03055,0.02681,0.01352,0.01454,0.003711,17.63,32.12,122.4,896.9,0.1525,0.6643,0.5539,0.2701,0.4264,0.1275
852783,"18,61,20,25,122.1,1894,0.0944,0.1066,0.149,0.07731,0.1697,0.05699,0.8529,1.849,5.632,93.54,0.01075,0.02722,0.05081,0.01911,0.02293,0.004217,21.31,27.26,139.9,1403,0.1338,0.1338,0.3446,0.149,0.2341,0.07421
852973,"15,3,25,27,102.4,731.4,0.1082,0.1697,0.1683,0.08751,0.1926,0.0654,0.439,1.012,4.498,43.5,0.005233,0.03057,0.03576,0.01083,0.01768,0.002967,20.27,36.71,149.3,1169,0.1641,0.611,0.6335,0.2034,0.4027,0.09076
853201,"17,57,15,05,115,95.1,0.09847,0.1157,0.09875,0.07953,0.1739,0.06149,0.6803,0.0225,4.655,61.1,0.005637,0.03033,0.03407,0.01354,0.01925,0.003742,20.81,19.52,134.9,1227,0.1255,0.2812,0.2489,0.1456,0.2756,0.07919
853401,"18,63,25,11,124.8,1880,0.1064,0.1087,0.2319,0.1244,0.2103,0.06197,0.8307,1.406,5.574,105,0.006248,0.03374,0.05196,0.01158,0.02007,0.00456,23.15,34.01,160.5,1670,0.1491,0.4257,0.6123,0.1048,0.3444,0.09782
853612,"11,84,18,7,77.93,448.6,0.1109,0.1516,0.1218,0.05182,0.2301,0.07799,0.4025,1.483,4.75,41,0.005531,0.02414,0.04205,0.01044,0.002773,0.005667,16.82,28,112,119.4,888.7,0.1637,0.5775,0.6996,0.1546,0.4761,0.1402
85382801,"8,17,82,23.58,112.8,899.3,0.1197,0.1496,0.2417,0.1203,0.2248,0.06302,0.6809,1.398,3.999,67.78,0.008268,0.03082,0.05042,0.01112,0.02102,0.003854,20.88,32.09,136.1,1344,0.1634,0.3559,0.5588,0.1947,0.353,0.08482
854002,"19,27,26,47,127.9,1162,0.09401,0.1719,0.1657,0.07593,0.1853,0.06261,0.5598,0.6862,3.528,68.17,0.005015,0.03318,0.03497,0.009643,0.01543,0.003896,24.15,30.9,161.4,1813,0.1509,0.609,0.6891,0.1785,0.3672,0.1123
854039,"16,13,17,88,107,807.2,0.104,0.1559,0.1354,0.07752,0.1990,0.06515,0.334,0.6857,2.183,35.03,0.004185,0.02688,0.02684,0.009067,0.01703,0.003817,20.21,27.16,132.7,1261,0.1446,0.5804,0.5274,0.1864,0.427,0.1233
```

Fig 3.4.1: Dataset Preview

2. The total number of the columns are 31.
3. Total number of rows are almost 570.
4. 80% of the data we used in our model to train.
5. 20% of the total data is used for testing purpose.
6. Dataset is saved in excel file which extension is .csv.

CHAPTER 4

Implementation and Algorithms

4.1 SVM (Support Vector Machine):

SVM can be a classifier of discrimination illustrated in a separating hyperplane. Much of the time, SVM can be a familiar regulated master learning law. It is used in classification problems often. Inside the SVM model, we can depict each object in n-dimensional field as a degree that would be worth each element in a particular coordinate (where n is many features). By Associate, the example of the 2 categories would be well understood. With the help of this estimate, we can understand the mechanism of the SVM classifier.

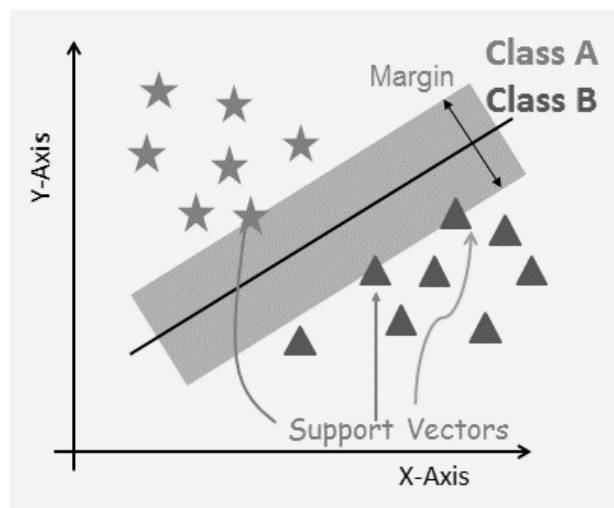


Figure 4.1.1: SVM figure

The key aim is the greatest feasible isolation of the data gathering. The function can be described as the margin between some of the closest data points. The target is to determine the maximum margin for a hyperplane between supportive vectors in the collection of data. In the steps outlined SVM looks for the nominal maximal hyperplane:

Build hyperplanes that better differentiate the groups. Figure on the left displaying three yellow, blue and orange hyperplanes. The higher classification of the blue and the orange is mistake, but the black accurately distinguishes the two classes.

Choose the proper hyperplane from each of the closest distance for complete separation.

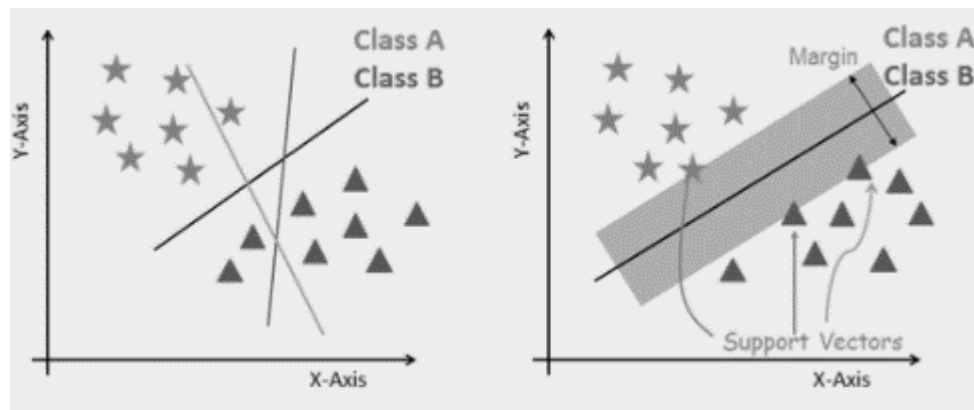


Figure 4.1.2: How SVM works

While Managing nonlinear and indistinguishable planes;

There are some complications regarding linear hyperplanes which cannot be overcome, as shown below (left-hand side).

SVM uses a kernel trick in that scenario to translate the establishment space into a higher dimensional feature space, as is shown at the right. On the x-axis and z-axis, the data points are shown ($z = x^2 = y^2$). Now, with linear segregation, you can conveniently differentiate these points.

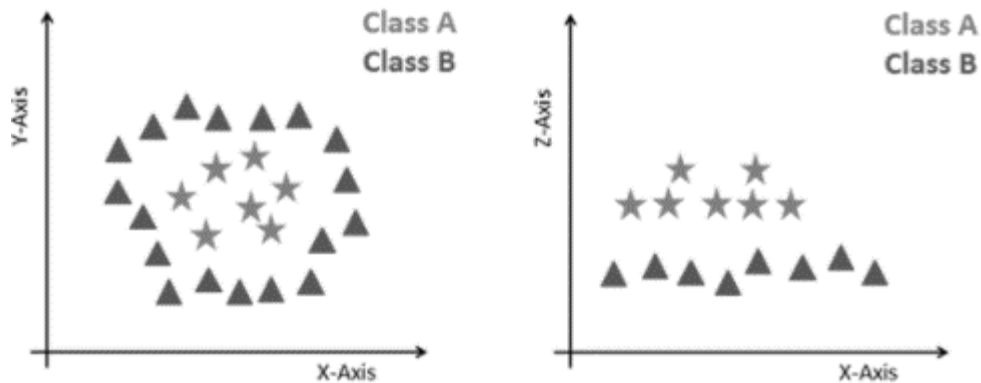


Figure 4.1.3: Managing nonlinear and indistinguishable planes

Kernel is used to execute the SVM algorithm in operation. A kernel can translate the data input space into the form necessary. SVM is using a kernel trick strategy. In this case, the kernel takes a small input space and renders it a larger space. In other words, it can be assumed that by introducing more dimensions, it transforms an unrelated problem into separable problems. In non-linear separation problems it is most useful. The kernel trick lets you construct a more detailed classification.

As a standard dot product, any two actions taken can be used as a linear kernel. The product between the two vectors is the sum within each pair of input data multiplying. The mathematical representation can be shown like this:

$$K(x, x_i) = \text{sum}(x * x_i)$$

The polynomial kernel would be a more widespread linear kernel structure. The kernel polynomial will discriminate input space, whether curved or nonlinear.

$$K(x, x_i) = 1 + \text{sum}(x * x_i)^d$$

Where d is the polynomial stage. $d=1$ is a practical implementation close to that. In the learning algorithm the degree must be manually set.

The radial functional kernel is a prominent kernel function widely used to help the classification of vector machines. In infinite dimensional space, RBF can remap input space.

$$K(x, x_i) = \exp(-\gamma * \sum ((x - x_i)^2))$$

Here gamma is a function varying between 0 and 1. A higher gamma value matches the training dataset, which induces unnecessary adaptation. Gamma=0.1 is known as a good predicted value. Throughout the learning algorithm the gamma value needs to be defined individually. Comparative to the Naive Bayes algorithm, SVM categories give reasonable precision and achieve faster prediction. They often use less memory and in decision-making, they use a sub-set of teaching points. SVM fits best with a strong separation margin and a wide range.

4.2 Random Forest Classifier:

Random forests are indeed an algorithm of supervised analysis. And can be used for regression along with classification. This algorithm might be the most versatile and easy to use. There are trees in a woodland. The more trees they have, the better a forest is. Random forests generate randomly chosen samples for decision-making trees, obtain precision from each tree and choose the best solution by votes. It also gives an excellent indication of the value of the function. Random forests have a wide range of uses, including recommendation systems, image recognition, and collection of functionalities. It is used for classifying faithful creditors, for detecting frauds and for anticipating diseases. The Boruta algorithm is based on which important characteristics in a data set are chosen.

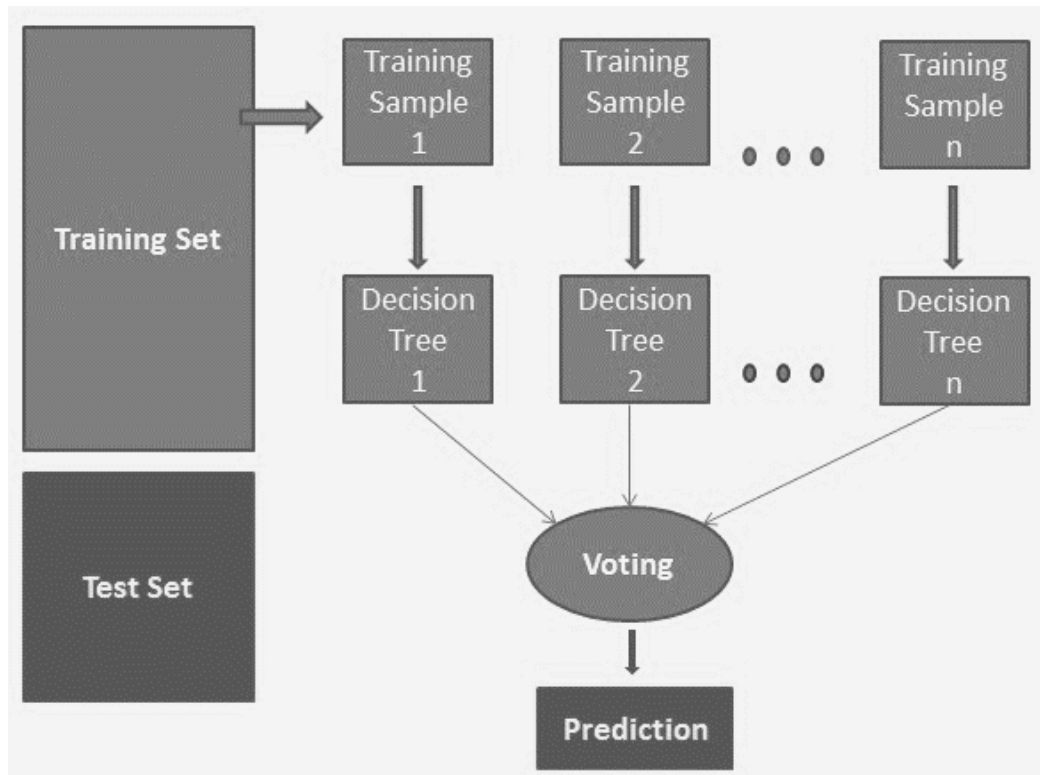


Figure 4.2.1: Steps of how Random forest classifier works

- 1) Chooses a random dataset sample.
- 2) Generate a decision tree and make a projection out of each decision tree for each sample. Carry out a referendum on each expected outcome.
- 3) Choose the outcome with the most votes as the overall projection.

As a consequence of the number of decision trees involved in the procedure, random forests are seen as an incredibly detailed and stable system. The matter of over-supply does not exist. The biggest explanation is that all forecasts take the average and exclude the predictions. The classification and regression questions can be used with both algorithms. The outliers can also be managed by random forest. In two manners, median values are used to override the continuous variables and the average weighted proximity of missing values is determined. Can get the relative value of the function, which helps to assign the classifier parameters. The random forest proposed technique implemented the general

bootstrap methodology for tree subjects. In the case of a training dataset $X=x_1 \dots x_n$ with $Y=y_1 \dots y_n$ responses, y_n boosting chooses the random sample and replaces the training set with the tree that represents the sampling samples. After processing, x' unseen sample projections can be accomplished by means of an average of x' for the whole tree structure on x' :

$$\hat{f} = \frac{1}{B} \sum_{b=1}^B f_b(x')$$

4.3 K Nearest Neighbor:

K Nearest Neighbor Classifier (KNN) is one of the most basic algorithms, easy to interpret, scalable and top-of-the-line. In the fields of economics, medicine, political science, handwriting recognition, visual recognition and video recognition, KNN is used.

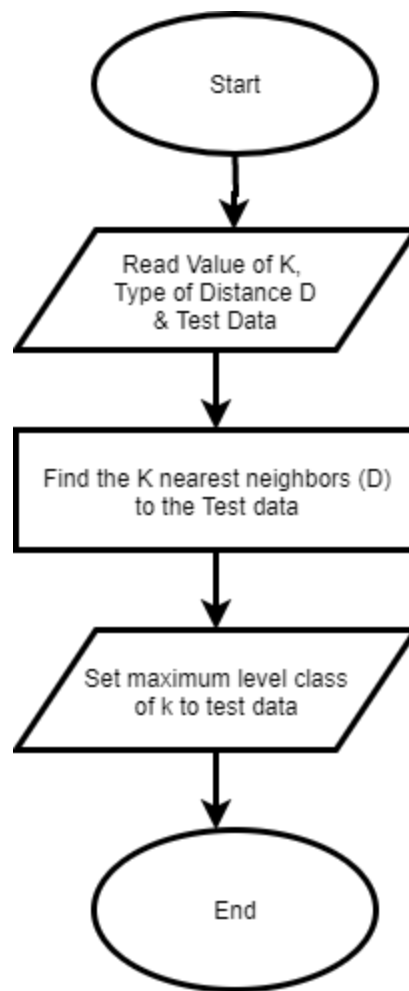


Figure 4.3.1: KNN flowchart

Euclidean metric equation:

$$d(x, x') = \sqrt{(x_1 - x'_1)^2 + \dots + (x_n - x'_n)^2}$$

Probability:

$$P(Y = j|X = x) = \frac{1}{K} \sum_{i \in A} I(y^{(i)} = j)$$

Financial firms can forecast consumer credit scores in credit ratings. Banking institutions forecast whether the loan is secure or dangerous in the disbursement of credit. Potential

candidates in two groups may vote or not vote in political science. KNN algorithm used in both regression and classification questions. KNN focused on an approach to feature resemblance. The learning algorithm is non-parametric and lazy. Non-parametric means for the corresponding distribution of data is not considered. The configuration of the model was described by the dataset. This is very useful in fact, if statistical scientific principles are not observed by most real-time datasets. Ignorant algorithm ensures that no training data centers are required to produce the model. All data used in the test process of preparation. That reduces the cost and pace of training and testing process. Checking time and memory means costly. In the worst case, more time would be taken to search all data points and to search all data points for information storage. The number of immediate neighbors is K in KNN. The biggest decisive factor is the number of neighbors. K is normally a strange number if the class numbers are 2. For $K=1$, the algorithm is referred to as the closest algorithm. This is the most straightforward scenario. Suppose $P1$ is the point to be expected by mark. Initially you've found the one point nearest to $P1$ as well as the label next to $P1$.

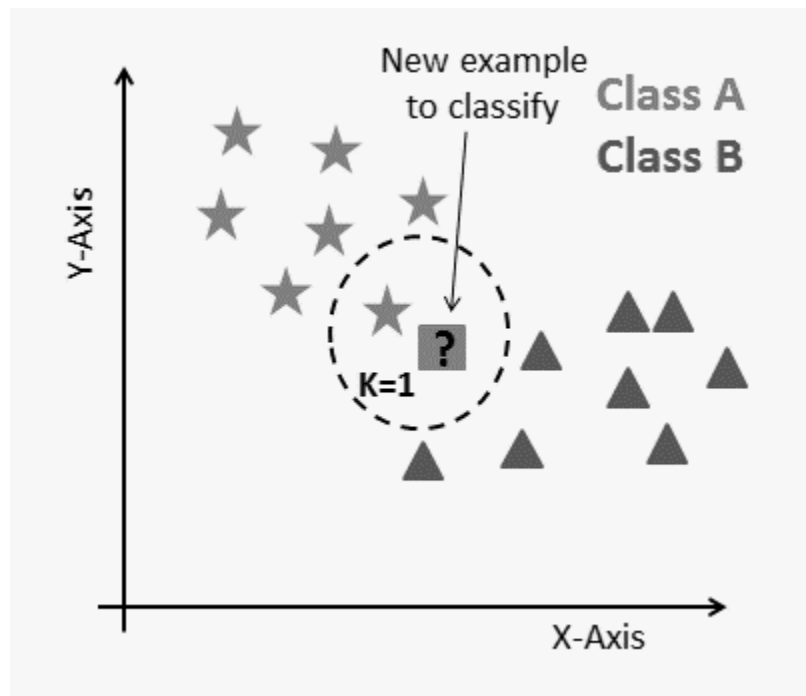


Figure 4.3.2: The process of KNN algorithm

P1 is the point to be projected by mark. Second, you will find the nearest k to P1 and then by plurality voting of their k neighbors you will classify points. The clause is taken for each item vote for its class and for the class with the most votes. The distance between points is found by means of distance measurements as Euclidean, Hamming and Manhattan as well as Minkowski as the nearest points are close. The following simple steps are taken by KNN:

- a) Range estimation
- b) Find the nearest vicinity
- c) Vote for markings

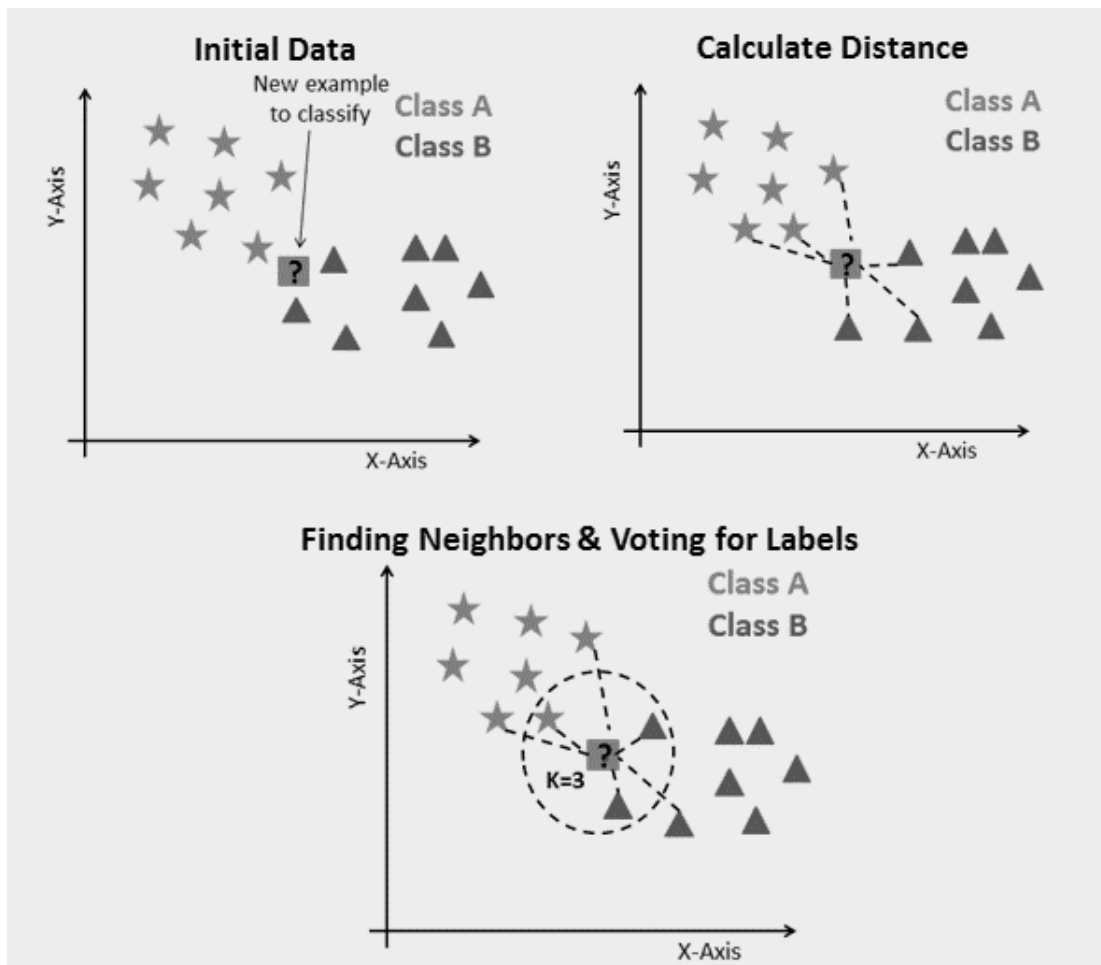


Figure 4.3.3: The process of KNN algorithm

K as a projection model regulating variable. There are no perfect neighbors for all data sets. Its own criteria for each dataset. For a limited number of neighbors, the vibration will have a larger impact on the results which will make computationally costly for a significant number of neighbors. Research has also shown that the smoother judgment limit with a small number of nodes is the most flexible match and has low bias but a high variance, and a large number of neighbors implies a lower variance but a greater bias. Generally, if the number of parties becomes even, data scientists just get an odd number.

$$d(p, q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \dots + (q_n - p_n)^2}$$

$$\sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$

4.4 Gaussian Naïve Bayes:

Naive Bayes is the simplest and most effective ranking algorithm for a wide variety of information. In a wide range of applications including spam processing, text classification, emotion interpretation and recommender mechanisms Naive Bayes classification is effectively used. It incorporates Bayes probability theorem for unspecified class prediction.

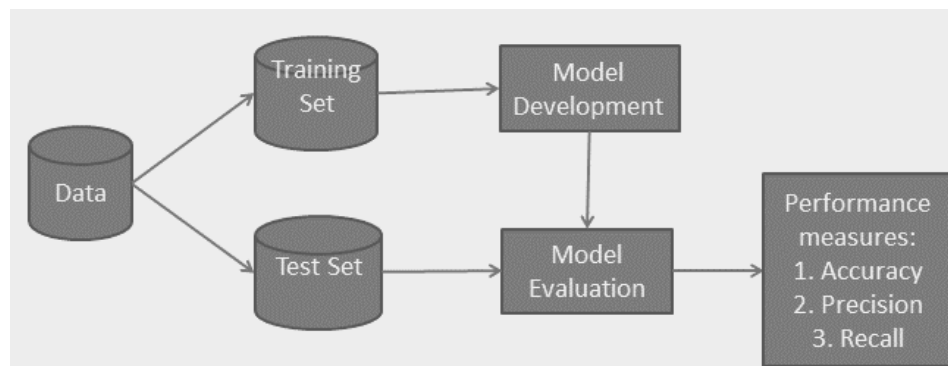


Figure 4.4.1: Naïve bayes working process

The division consists of two stages, a study and the appraisal process. The classifier trains

its model on a given data set during the learning process and checks its classifier output during the assessment phase. Evaluated results on different metrics including exactness, mistake, precision and reminder. Naive Bayes is a Bayes theorem-based mathematical classification method. This algorithm is one of the simplest supervised. The fast, accurate and reliable algorithm is Naive Bayes. High precision and speed classifiers of Naive Bayes are possible on large datasets.

Phase 1: Pre-value estimates with some class marks

Phase 2: Find the chance of possibility for each class attribute

Phase 3: In Bayes Formula, put this value and determine the likelihood for later.

Phase 4: See which class is more probable since the feedback is in the higher likelihood class.

The classification scheme of Naive Bayes suggests that the influence of a given function in a class is independent from other features. For instance, a loan applicant is suitable or not based on wages, past experience, age and place of credit and transaction. And if these traits are interdependent, they are nevertheless treated separately. This presumption makes programming easy and is thus assumed to be naive. This presumption is referred to as class emancipation.

$$P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)}$$

$P(h)$: the possibility of hypothesis h (regardless of the data). This is referred to as the previous possibility of h .

$P(D)$: the data chance (regardless of the hypothesis). This is referred to as the previous opportunity.

$P(h|D)$: hypothesis high probability h given the D data. This is considered a posterior likelihood.

$P(D|h)$: the mere possibility of data d because the hypothesis h is valid. This is considered a posterior likelihood.

4.5 Logistics Regression:

LR is a mathematical technique for binary class prediction. The consequence or objective variable is dichotomous in nature. This means that only two classes are possible. TRUE=1 or FALSE=0 respectively. Regression of logistics is much like linear regression. As a general linear model, we can say that. Equation of linear regression.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

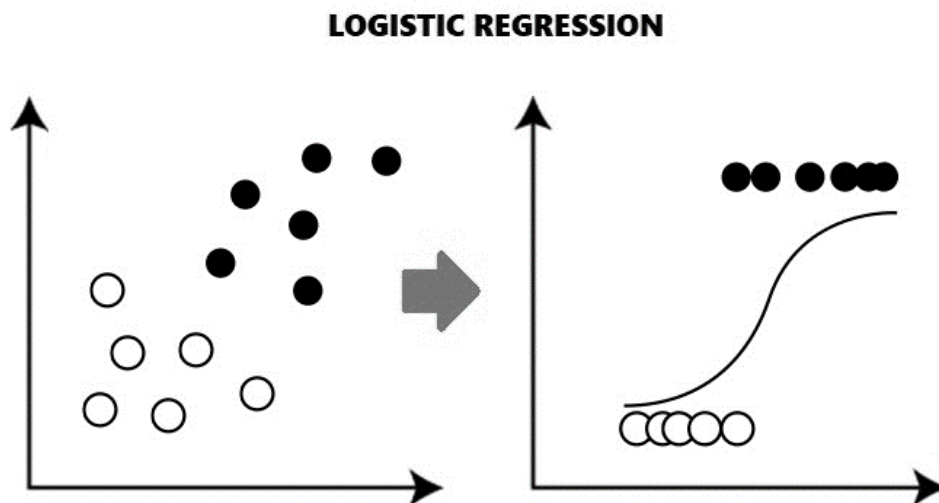


Figure 4.5.1: Logistic Regression

Sigmoid function:
$$p = \frac{1}{1+e^{-y}}$$

Final equation after applying sigmoid function:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$

4.6 Decision Tree:

A decision tree is a flux-chart tree structure in which a method (or attribute) is represented by the internal node, the branch is a law of decision and each leaf node represents the consequence. The top node is known as the root node in a decision tree. It is based on the value of the attribute. It recursively divides the tree into periodic partitions. This fluid diagram form lets you make your choices. It is a display like a diagram that conveniently imitates the reasoning of humans. That's why decision-making trees are easy to understand.

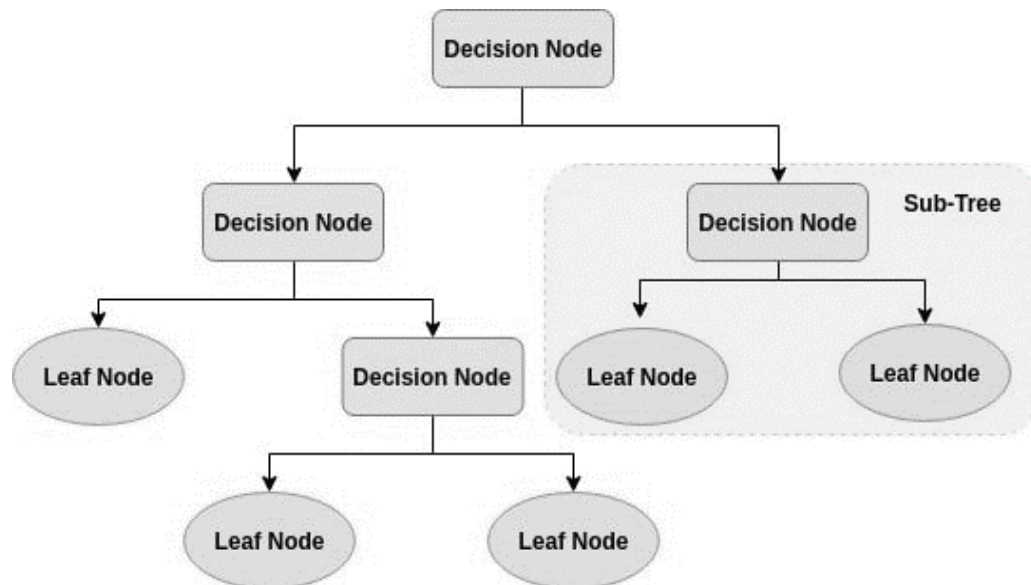


Figure 4.6.1: Decision tree

Decision Tree is an ML algorithm style white box. It has an intrinsic decision-making logic, and the algorithms such as neural network do not contain any of the black box form. Its training time is shorter than the algorithm for the neural network. Significant impact on the number of records and number of features in the data provided, time complexity of decision tree. The decision tree is a non-parametric or variance free process, which does not rely on expectations of likelihood allocation. Decision trees can handle with good accuracy high-dimensional results. Select the best characteristic to separate the documents while using

the ASM. Make a judgment node attribute and split the data collection into thread. Starts the construction of a tree by repeating this step recursively for each child until one condition matches: Both tuples are of the same value as attributes. No more characteristics are left. No more incidents.

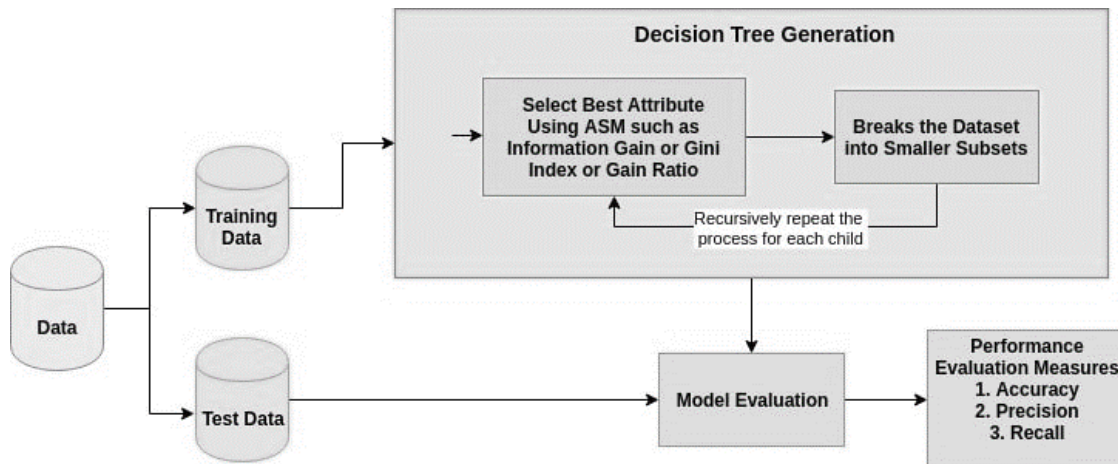


Figure 4.6.2: Working process of Decision Tree

Attribute Selection Measure:

The selection measure of the characteristics is a heuristic approach for choosing the partitioning criteria in the best way possible. The splitting rules are also known as it aids in the detection of splitting points for tuples at a specific node. By describing the given dataset, ASM lists a rank for each feature(s). The best score is chosen as a dividing characteristic. For an underway attribute, split points for branches must also be established. Knowledge benefit, gain ratio and Gini index are the most commonly used performance behavior.

Information Gain:

The entropy term that measures the input impurity developed. Entropy, called randomness or impurity in the device, is used in physics and mathematics. It corresponds to impurity in a group of examples of information theory. The decline in entropy is the gain in information. The data benefit measures the discrepancy between the entropy and the average entropy after division of the dataset based on the values of the attribute. The decision tree algorithm ID3 (Iterative Dichotomies) uses the benefit of knowledge.

$$Info(D) = - \sum_{i=1}^m P^i \log_2 p^i$$

$$Info_A(D) = \sum_{j=1}^v \frac{|D_j|}{|D|} \times Info(D_j)$$

$$Gain(A) = Info(D) - Info_A(D)$$

$$\Delta Gini(A) = Gini(D) - Gini_a(D)$$

CHAPTER 5

Experimental Results, Discussion and Conclusion

5.1 Introduction

In this section, we will talk about the experimental result. We have only discussed all six algorithms which we will use. So now we will see how much accurately those algorithms worked. And we will compare the accuracy with all six algorithms.

Steps that we follow to complete this research work are:

Step-1: Collect dataset

Step-2: Dataset pre-processing

Step-3: Import different libraries and other necessary things

Step-4: Split our dataset

Step-5: Create models for all 6 algorithms

Step-6: Train with all the 6 different machine learning algorithms

Step-7: Find the accuracy of all the algorithms

Step-8: Compare which algorithm gives the best result

These are the steps we follow to complete this research work.

5.2 Experimental Results

We know no machine can give us 100% output. Similarly, we can train our model and change some parameters to get more accuracy and train properly. But the accuracy we get from different algorithms are pretty good.

There are some photos given below which is showing our research work in a short form. In these photos we are showing precision, recall, f1 score, support, accuracy, heatmap.

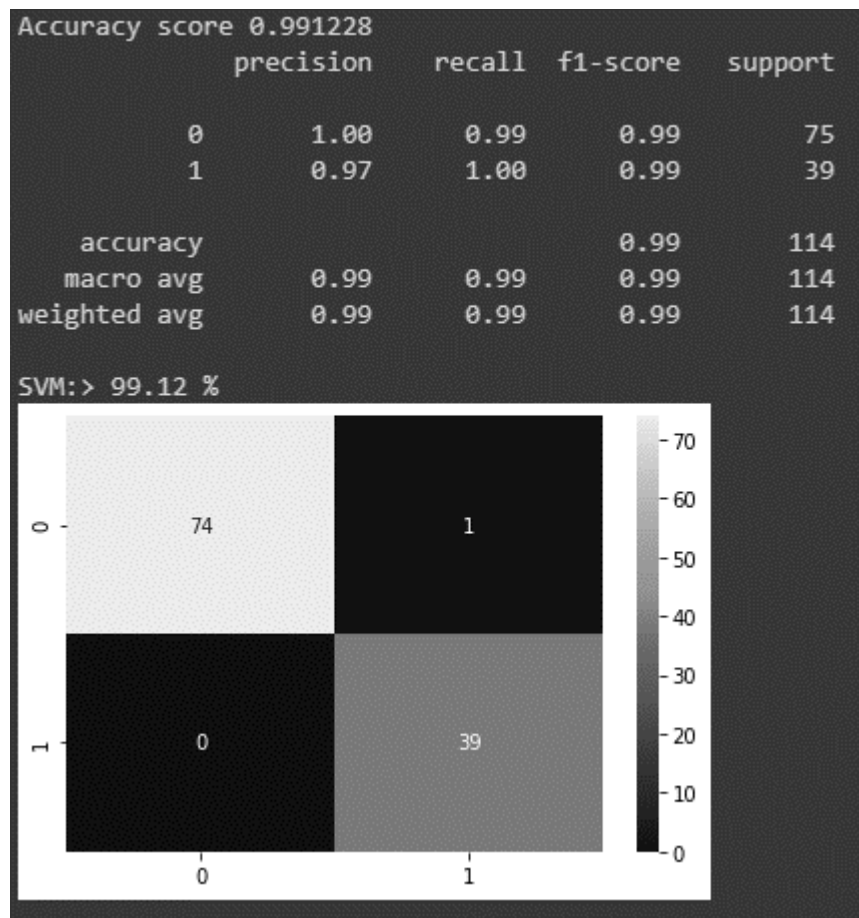


Figure 5.2.1: SVM Result

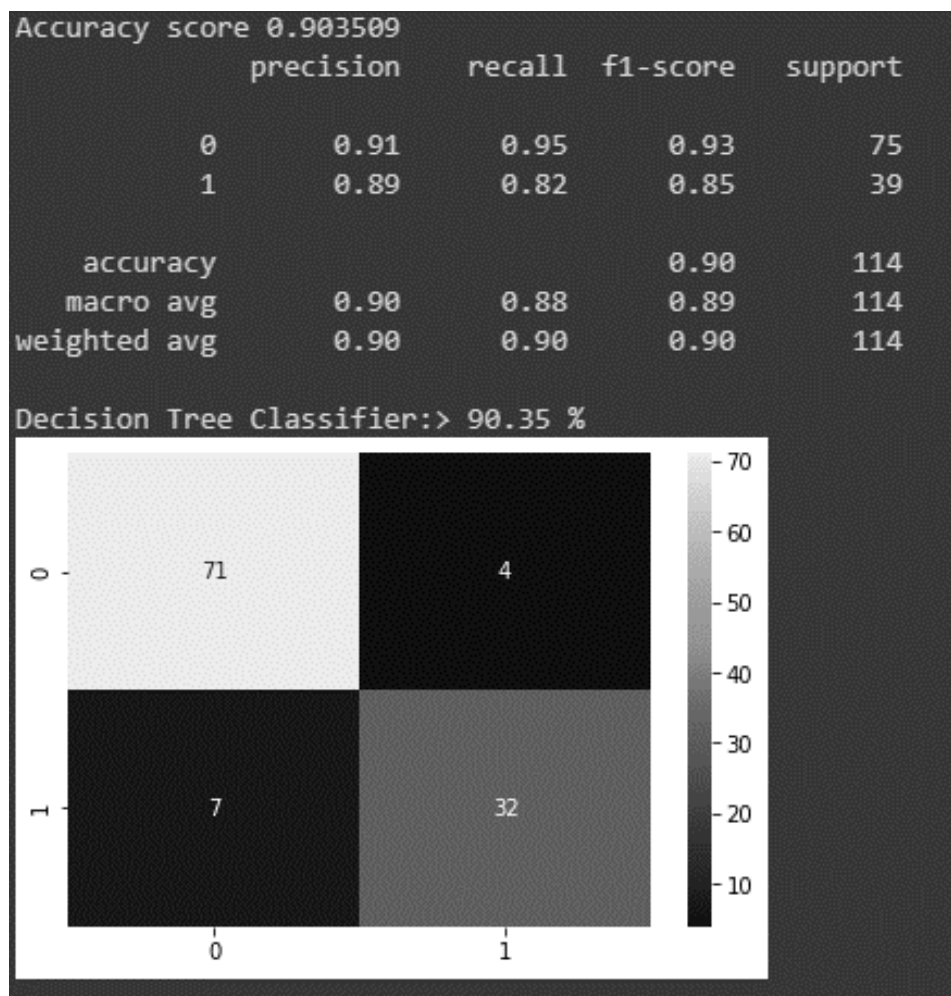


Figure 5.2.2: Decision tree Result

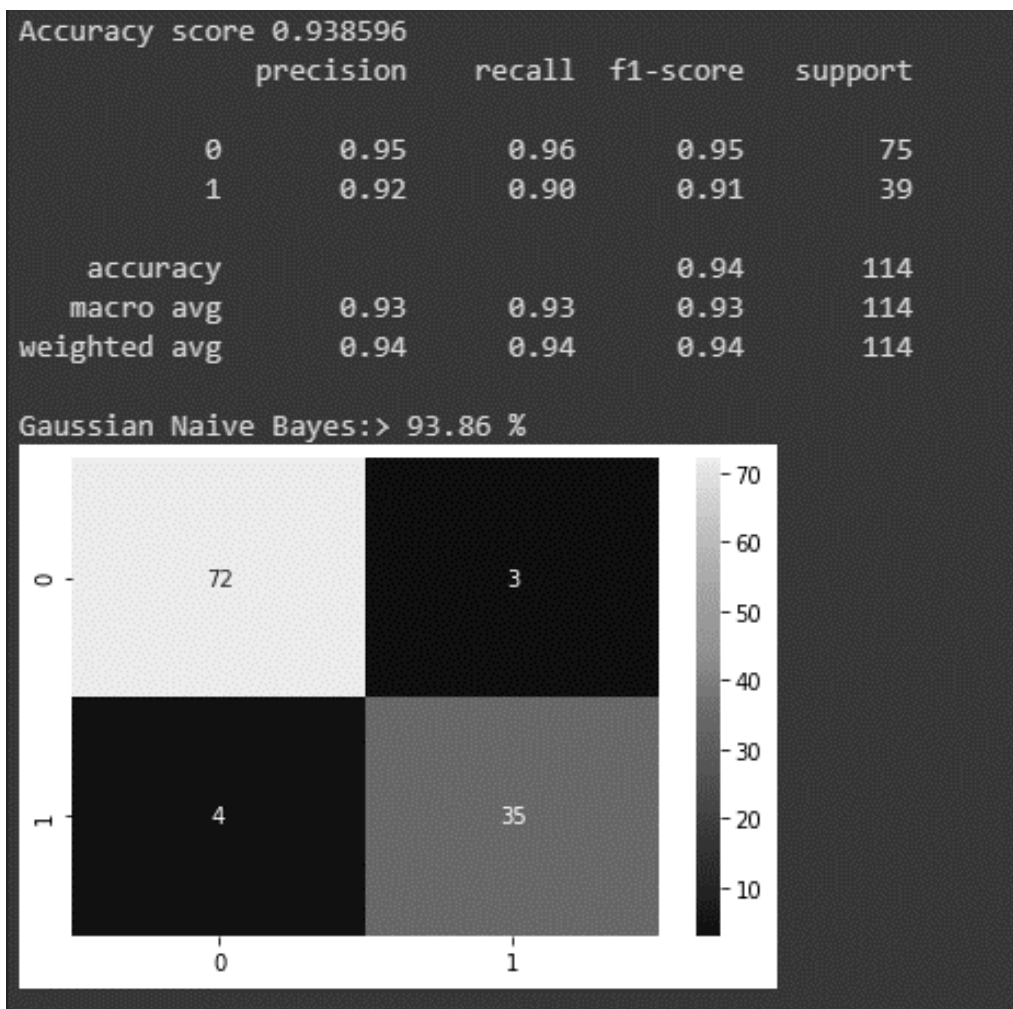


Figure 5.2.3: Gaussian Naïve Bayes Result

```

Accuracy score 0.982456
              precision    recall  f1-score   support

      0       0.97       1.00       0.99        75
      1       1.00       0.95       0.97        39

 accuracy          0.98        114
 macro avg         0.99        114
 weighted avg      0.98        114

```

K-Neighbors Classifier:> 98.25 %

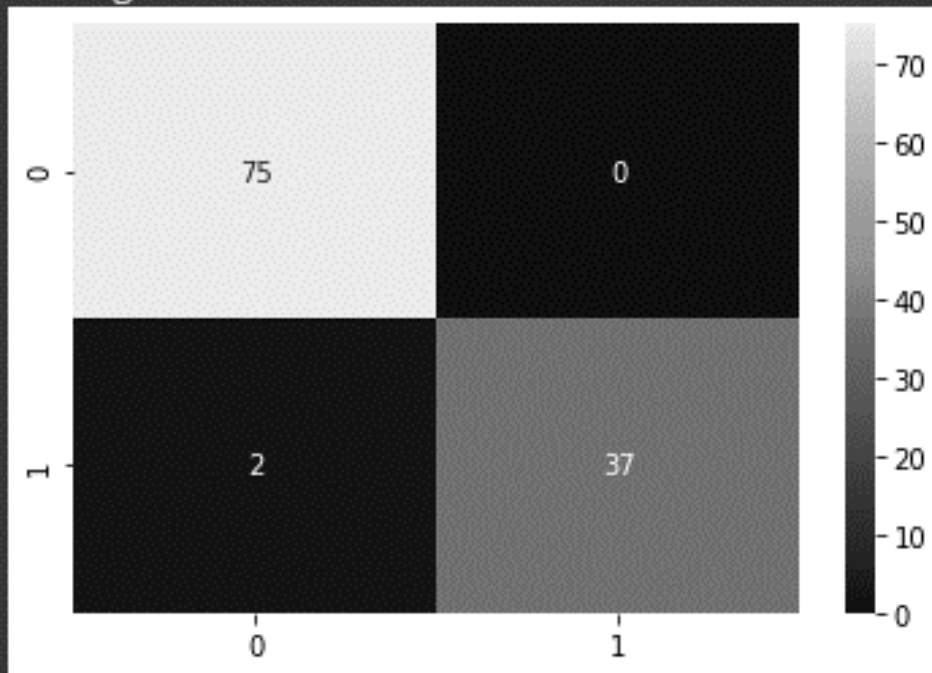


Figure 5.2.4: KNN Result

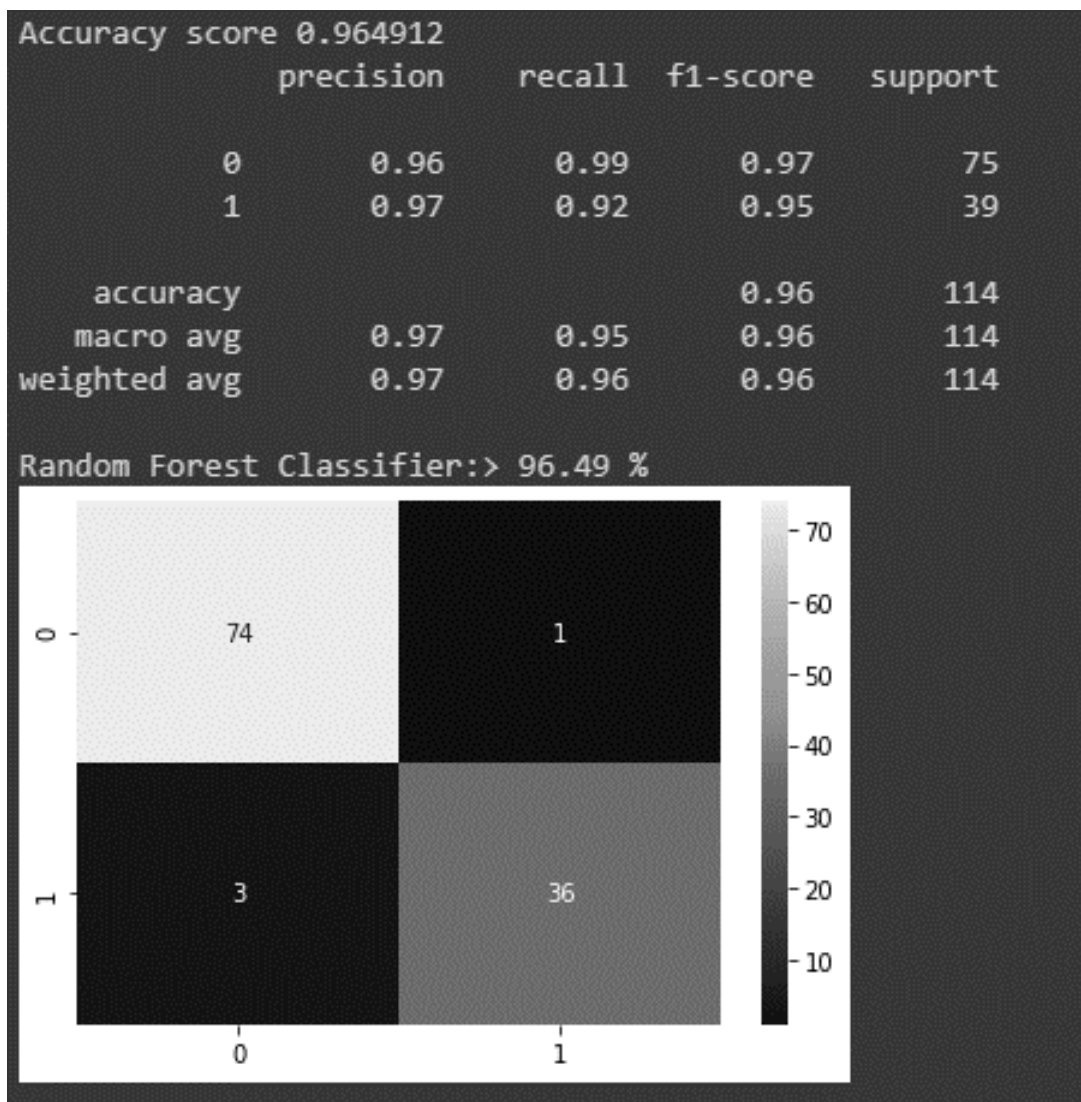


Figure 5.2.5: Random forest Result

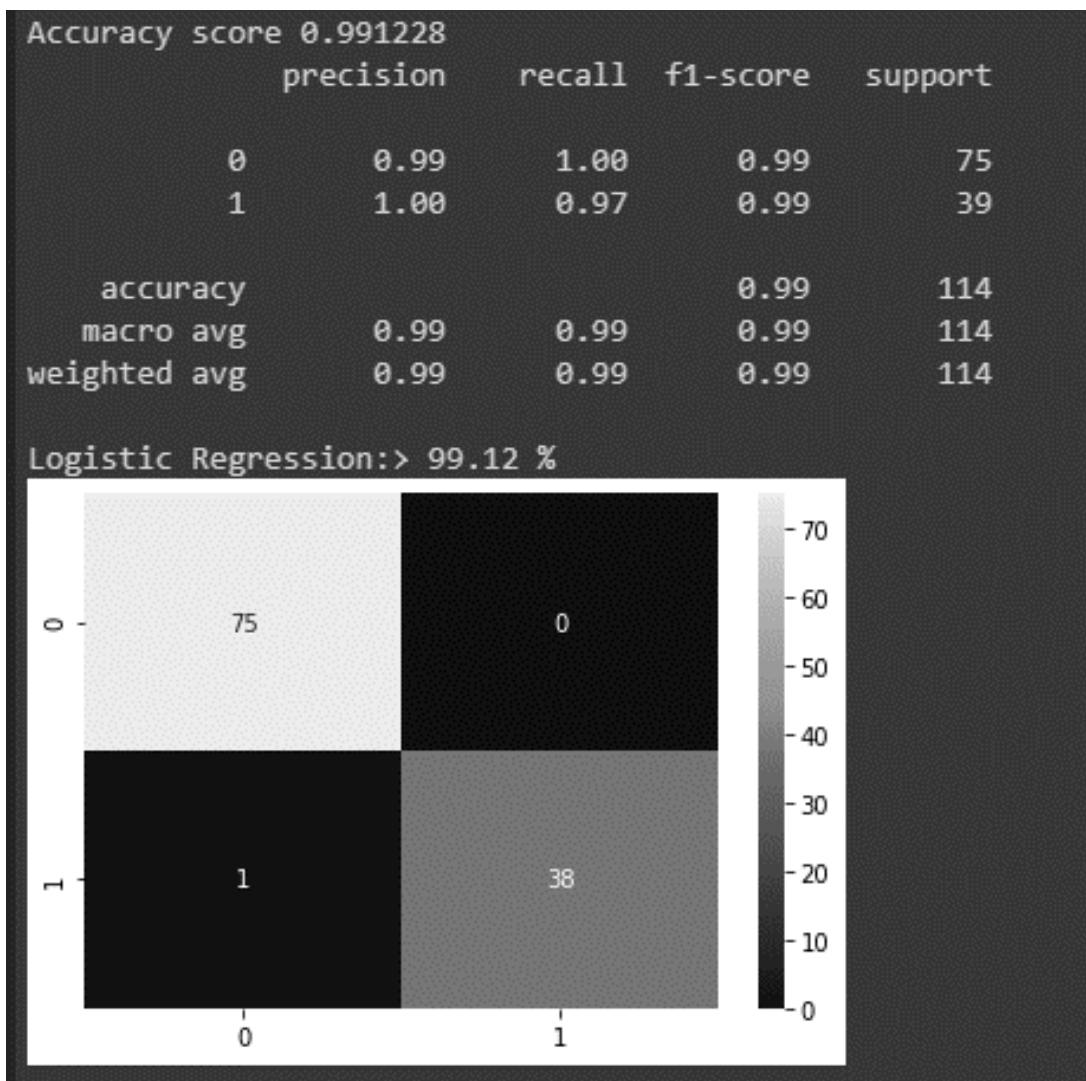


Figure 5.2.6: Logistic Regression Result

TABLE 1: RESULTS OF DIFFERENT ALGORITHMS

Algorithm name	Precision		Recall		F1-score		Accuracy (%)
	0	1	0	1	0	1	
SVM	1.00	0.97	0.99	1.00	0.99	0.99	99.12
Decision Tree	0.91	0.89	0.95	0.82	0.93	0.85	90.35
Gaussian Naïve Bayes	0.95	0.92	0.96	0.90	0.95	0.91	93.86
KNN	0.97	1.00	1.00	0.95	0.99	0.97	98.25
Random Forest	0.96	0.97	0.99	0.92	0.97	0.95	96.49
Logistics Regression	0.99	1.00	1.00	0.97	0.99	0.99	99.12

We can see the accuracy of 6 different algorithms. All the algorithms are showing above 90% accuracy. But SVM and Logistic Regression these two algorithms are showing 99.12% accuracy. So, if we compare all 6 algorithms, we can say that SVM and Logistic Regression can predict or detect the breast cancer most accurately.

There some other figures we can include to give some more idea about our research work.

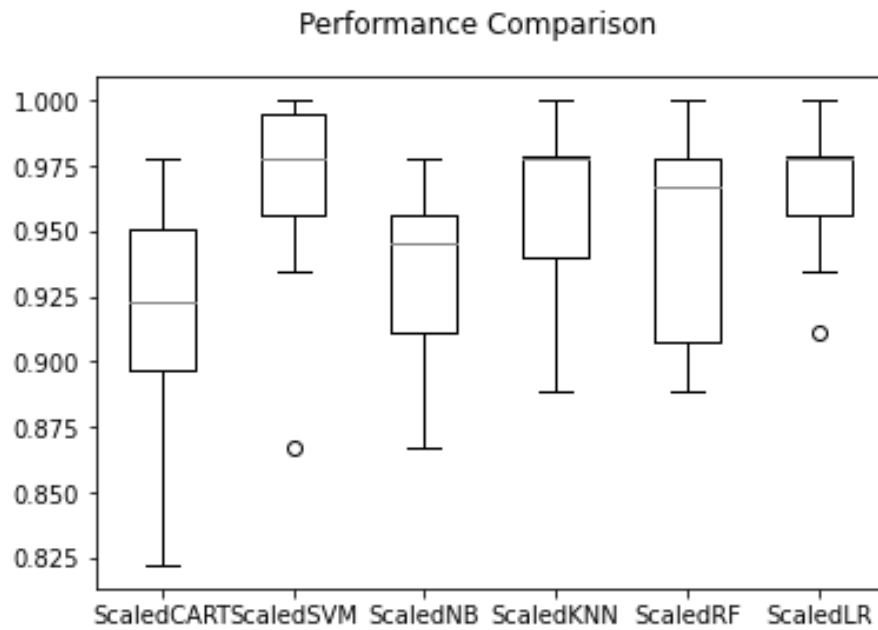


Figure 5.2.7: Performance comparison

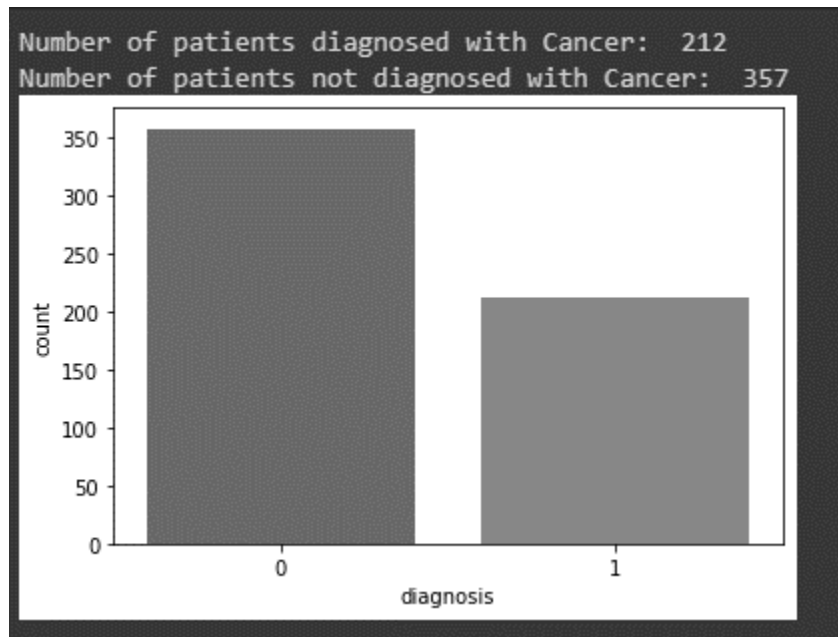


Figure 5.2.8: Number of Benign and Malignant cases from the dataset

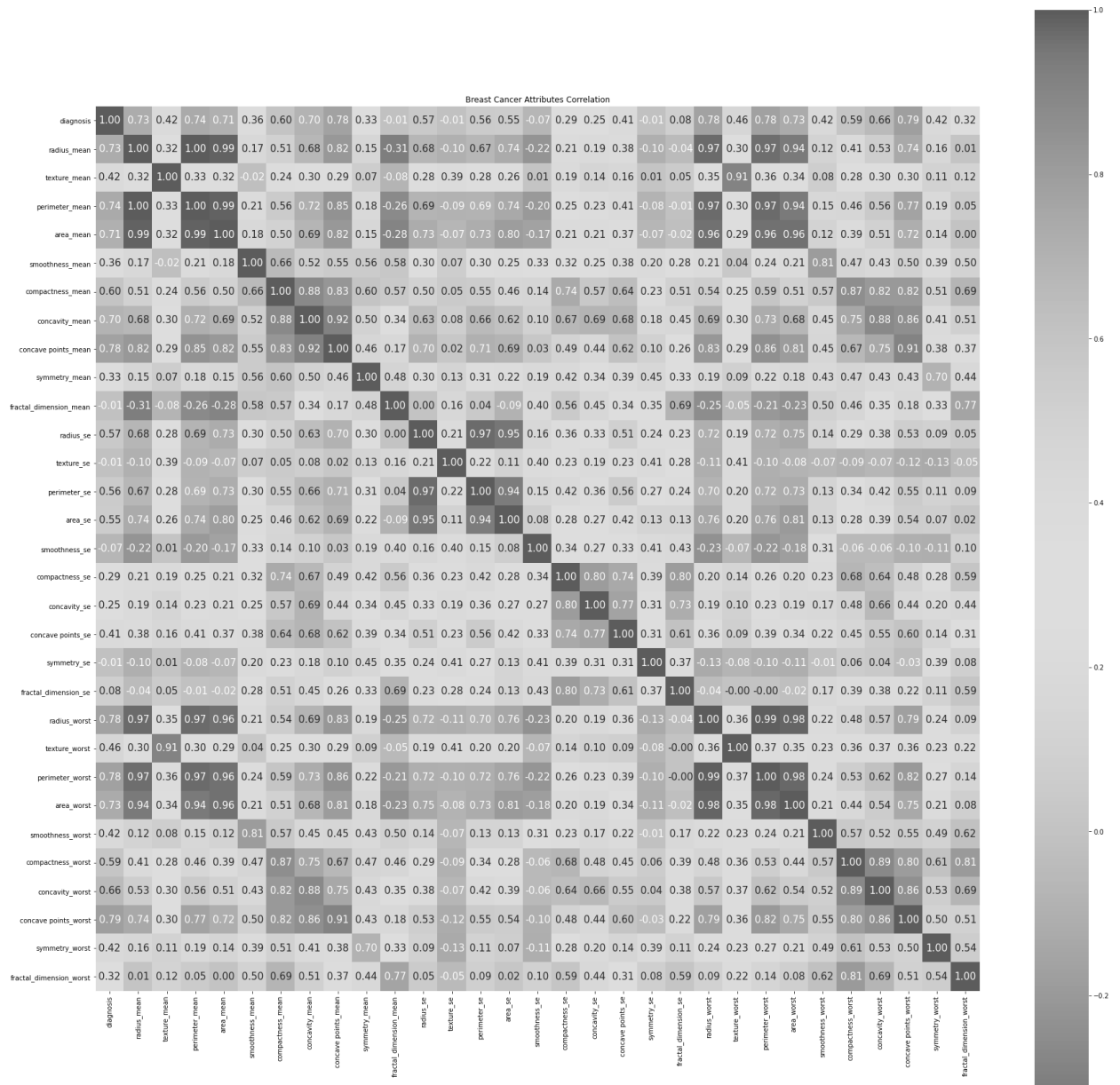


Figure 5.2.9: Breast Cancer Attributes Correlation Heatmap

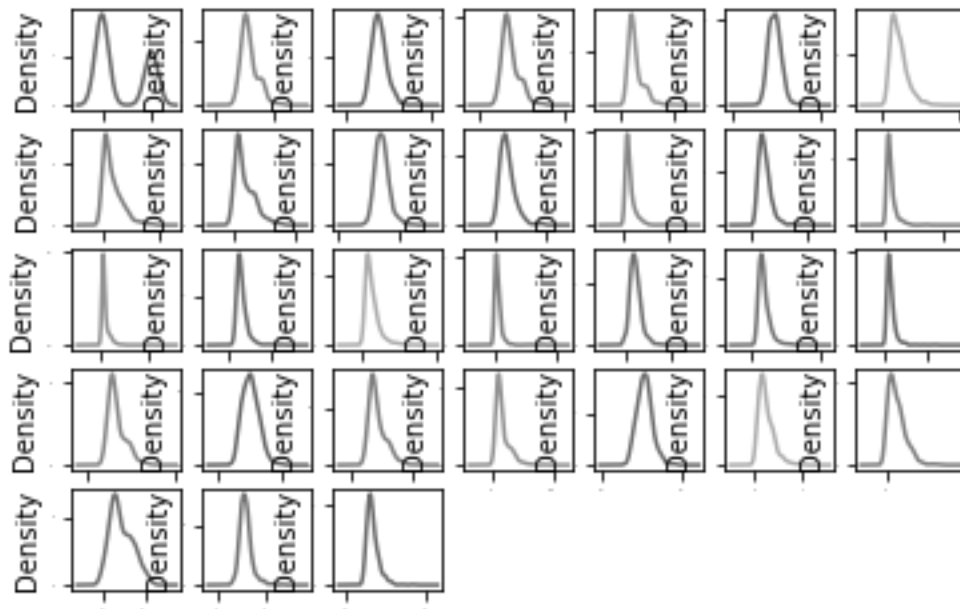


Figure 5.2.10: General Gaussian Distribution

5.3 Conclusion

We can conclude our research work by saying that we have tried to use different algorithms to compare the accuracy of detecting breast cancer. We get the highest accuracy from 2 algorithms which are support vector machine or SVM and another one is logistic regression algorithm. As we have worked with only 569 data, we have planned to increase our data and make the dataset bigger with more accurate values. We have the plan to use other machine learning algorithms to compare that if any other algorithm can give us accuracies more than close enough like support vector machine and Logistic regression algorithms.

APPENDICES

Annex refers to the representation of the analysis. We faced too many questions to complete the project, one of which was to settle on the methodology of our work. Another concern was data collection, this was an immense challenge. Yet we are done by this challenge and complete our job successfully. Breast cancer identification and identification is the key function of our application. Our application helps users to quickly understand and determine the signs. And the number of patients is therefore limited. Therefore, we prefer to conclude that our program "**Performance Analysis of Classification Algorithm for Breast Cancer Prediction**" would be incredibly constructive and effective for consumers.

REFERENCES

- [1] Hiba Asri, H. M. (2016). Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. *Procedia Computer Science*, 83, 1064-1069. doi:<https://doi.org/10.1016/j.procs.2016.04.224>.
- [2] Ganggayah, M.D., Taib, N.A., Har, Y.C. *et al.* Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC Med Inform Decis Mak* **19**, 48 (2019).
- [3] M. Amrane, S. Oukid, I. Gagaoua and T. Ensarlı, "Breast cancer classification using machine learning," 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), Istanbul, 2018, pp. 1-4, DOI: 10.1109/EBBT.2018.8391453.
- [4] S. Sharma, A. Aggarwal, and T. Choudhury, "Breast Cancer Detection Using Machine Learning Algorithms," 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), Belgaum, India, 2018, pp. 114-118, DOI: 10.1109/CTEMS.2018.8769187.
- [5] Y. A. Hamad, K. Simonov, and M. B. Naeem, "Breast Cancer Detection and Classification Using Artificial Neural Networks," 2018 1st Annual International Conference on Information and Sciences (AiCIS), Fallujah, Iraq, 2018, pp. 51-57, DOI: 10.1109/AiCIS.2018.00022.
- [6] H. Jouni, M. Issa, A. Harb, G. Jacquemod and Y. Leduc, "Neural Network architecture for breast cancer detection and classification," 2016 IEEE International Multidisciplinary Conference on Engineering Technology (IMCET), Beirut, 2016, pp. 37-41, DOI: 10.1109/IMCET.2016.7777423.
- [7] C. Shahnaz, J. Hossain, S. A. Fattah, S. Ghosh and A. I. Khan, "Efficient approaches for accuracy improvement of breast cancer classification using Wisconsin database," 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), Dhaka, 2017, pp. 792-797, DOI: 10.1109/R10-HTC.2017.8289075.
- [8] D. S. Jacob, R. Viswan, V. Manju, L. PadmaSuresh and S. Raj, "A Survey on Breast Cancer Prediction Using Data Mining Techniques," 2018 Conference on Emerging Devices and Smart Systems (ICEDSS), Tiruchengode, 2018, pp. 256-258, DOI: 10.1109/ICEDSS.2018.8544268.
- [9] N. Fatima, L. Liu, S. Hong, and H. Ahmed, "Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis," in *IEEE Access*, vol. 8, pp. 150360-150376, 2020, DOI: 10.1109/ACCESS.2020.3016715.
- [10] S. Kabiraj et al., "Breast Cancer Risk Prediction using XGBoost and Random Forest Algorithm," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2020, pp. 1-4, DOI: 10.1109/ICCCNT49239.2020.9225451.

- [11] T. Padhi and P. Kumar, "Breast Cancer Analysis Using WEKA," 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2019, pp. 229-232, DOI: 10.1109/CONFLUENCE.2019.8776911.
- [12] A. M. Ibraheem, K. H. Rahouma, and H. F. A. Hamed, "Automatic MRI Breast tumor Detection using Discrete Wavelet Transform and Support Vector Machines," 2019 Novel Intelligent and Leading Emerging Sciences Conference (NILES), Giza, Egypt, 2019, pp. 88-91, DOI: 10.1109/NILES.2019.8909345.
- [13] M. Amrane, S. Oukid, I. Gagaoua, and T. Ensari, "Breast cancer classification using machine learning," 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), Istanbul, 2018, pp. 1-4, DOI: 10.1109/EBBT.2018.8391453.
- [14] N. Priya and G. Shobana, "Potential Breast Cancer Drug Prediction using Machine Learning Models," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore, India, 2020, pp. 1-6, DOI: 10.1109/ic-ETITE47903.2020.288.
- [15] A. Ivaturi, A. Singh, B. Gunanvitha and K. S. Chethan, "Soft Classification Techniques for Breast Cancer Detection and Classification," 2020 International Conference on Intelligent Engineering and Management (ICIEM), London, United Kingdom, 2020, pp. 437-442, DOI: 10.1109/ICIEM48762.2020.9160219.
- [16] A. K. Das, S. K. Biswas, A. Mandal, and M. Chakraborty, "A Neural Expert System to Identify Major Risk Factors of Breast Cancer," 2020 IEEE International Conference for Innovation in Technology (INOCON), BANGLURU, 2020, pp. 1-4, DOI: 10.1109/INOCON50539.2020.9298261.
- [17] Ali Al Bataineh, "A Comparative Analysis of Nonlinear Machine Learning Algorithms for Breast Cancer Detection," *International Journal of Machine Learning and Computing* vol. 9, no. 3, pp. 248-254, 2019.
- [18] ICBRA '18: 2018 5th International Conference on Bioinformatics Research and Applications Hong Kong December 2018, Association for Computing Machinery New York NY United States, ISBN: 978-1-4503-6611-3
- [19] Prediction of benign and malignant breast cancer using data mining techniques. Volume: 12 issue: 2, page(s): 119-126. Article first published online: February 20, 2018; Issue published: June 1, 2018. Received: September 30, 2017; Accepted: January 04, 2018
- [20] Juneja, K., Rana, C. An improved weighted decision tree approach for breast cancer prediction. *Int. j. inf. tecnol.* 12, 797–804 (2020). <https://doi.org/10.1007/s41870-018-0184-2>

- [21] Yue W, Wang Z, Chen H, Payne A, Liu X. Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis. *Designs*. 2018; 2(2):13. <https://doi.org/10.3390/designs2020013>
- [22] Hiba Asri, Hajar Mousannif, Hassan Al Moatassime, Thomas Noel, Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis, *Procedia Computer Science*, Volume 83, 2016, Pages 1064-1069, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2016.04.224>. (<https://www.sciencedirect.com/science/article/pii/S1877050916302575>), Keywords: Breast cancer; SVM; NB; C4.5; k-NN; Classification; Efficiency; Effectiveness.
- [23] Islam, M.M., Haque, M.R., Iqbal, H. *et al.* Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques. *SN COMPUT. SCI.* **1**, 290 (2020). <https://doi.org/10.1007/s42979-020-00305-w>
- [24] S. N. Singh and S. Thakral, "Using Data Mining Tools for Breast Cancer Prediction and Analysis," 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2018, pp. 1-4, doi: 10.1109/CCAA.2018.8777713.
- [25] A. Bharat, N. Pooja and R. A. Reddy, "Machine Learning algorithms for breast cancer risk prediction and diagnosis," 2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C), Bangalore, India, 2018, pp. 1-4, doi: 10.1109/CIMCA.2018.8739696.
- [26] M. R. Al-Hadidi, A. Alarabeyyat and M. Alhanahnah, "Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm," 2016 9th International Conference on Developments in eSystems Engineering (DeSE), Liverpool, 2016, pp. 35-39, doi: 10.1109/DeSE.2016.8.
- [27] M. I. H. Showrov, M. T. Islam, M. D. Hossain and M. S. Ahmed, "Performance Comparison of Three Classifiers for the Classification of Breast Cancer Dataset," 2019 4th International Conference on Electrical Information and Communication Technology (EICT), Khulna, Bangladesh, 2019, pp. 1-5, doi: 10.1109/EICT48899.2019.9068816.
- [28] S. Das and D. Biswas, "Prediction of Breast Cancer Using Ensemble Learning," 2019 5th International Conference on Advances in Electrical Engineering (ICAEE), Dhaka, Bangladesh, 2019, pp. 804-808, doi: 10.1109/ICAEE48663.2019.8975544.
- [29] Y. Khoudfifi and M. Bahaj, "Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification," 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), Kenitra, 2018, pp. 1-5, doi: 10.1109/ICECOCS.2018.8610632.

PLAGIARISM

Final Test

ORIGINALITY REPORT

23%
SIMILARITY INDEX

15%
INTERNET SOURCES

6%
PUBLICATIONS

16%
STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	6%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	5%
3	Submitted to Rochester Institute of Technology Student Paper	1%
4	www.datacamp.com Internet Source	1%
5	mmburublog.wordpress.com Internet Source	1%
6	Submitted to Arab Open University Student Paper	1%
7	Submitted to National College of Ireland Student Paper	1%
8	Submitted to Griffith College Dublin Student Paper	1%
9	ieeexplore.ieee.org Internet Source	1%