APPLYING ML ALGORITHMS FOR PREDICTING BREAST CANCER

BY

SAIFA SABRINA MIM
ID: 213-25-041

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

Supervised By

Professor Dr. Md. Fokhray Hossain
Professor
Department of CSE
Faculty of Science and Information Technology
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2023

APPROVAL

This Thesis titled **"Applying ML Algorithms for Predicting Breast Cancer"**, submitted by Saifa Sabrina Mim, ID No: 213-25-041 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 M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 17-01-2023.

BOARD OF EXAMINERS

17/01/23

Dr. Touhid Bhuiyan, PhD
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman

Ms. Nazmun Nessa Moon
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

ফিজার আহমেদ

Dr. Fizar Ahmed
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Safaet

Md. Safaet Hossain
Associate Professor & Head
Department of Computer Science and Engineering
City University

External Examiner

CS CamScanner

DECLARATION

We hereby declare that, this thesis has been done by us under the supervision of **Professor Dr. Md. Fokhray Hossain, 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:

Hossain

Professor Dr. Md. Fokhray Hossain
Professor
Department of CSE
Daffodil International University

Submitted by:

Saifa Sabrina Mim
ID: 213-25-041
Department of CSE
Daffodil International University

CS CamScanner

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty **ALLAH** for His divine blessing makes us possible to complete the thesis successfully.

We really grateful and wish our profound our indebtedness to **Professor Dr. Md. Fokhray Hossain, Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Data Mining*" to carry out this thesis. His endless patience ,scholarly guidance ,continual encouragement , constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Professor Dr. Touhid Bhuiyan**, Head**,** Department of CSE, for his kind help to finish my thesis 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

Cancer is, without a doubt, one of the worst illnesses in human history. Data from throughout the globe shows that we now face a number of serious health challenges. There is a significant risk of fatalities among cancer patients, and this is especially true when we examine instances of people with more than one kind of the disease. These symptoms are indicative of an early stage of cancer and the patient is currently at that stage. If a breast cancer tumor can be found at an early stage and its control functions can be suppressed to prevent it from growing or spreading to other organs, the survival rate for people with breast cancer may be increased. A better prognosis is possible. Moreover, if we consider the progression of cancer through time, we can observe that breast cancer is now one of the most prevalent forms of the disease. If we do our research, we'll find that there are many different kinds of data about breast cancer.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Background of The Research

As one of the deadliest and most frequent cancers in women today, breast cancer is a global health crisis. If we look at the numbers from across the world, we can see that there are a lot of serious problems with regards to health right now. It is clear from a review of cancer statistics that the mortality rate for cancer patients is substantial. There are recognizable signs that cancer has reached an early stage in a patient's body. [10] Breast cancer survival rates may be improved via early detection of tumors, aggressive treatment of those tumors, and careful monitoring to ensure that the disease does not grow or spread to other parts of the body. [11] Furthermore, if we examine the development of various phases of various malignancies, we can observe that breast cancer is one of the most prevalent for the current condition. A thorough examination reveals many types of breast cancer data collections. For the purpose of diagnosing and categorizing breast cancer, several researchers use a wide variety of approaches. Though there is a great deal of study on this subject, most authors only used a few different algorithms at most. We're working to combine the findings of many machine learning algorithms and enhance the number of algorithms used in our study. Here, we apply machine learning to the problem of detecting and diagnosing breast cancer. Finding the greatest candidate for breast cancer treatment early in the process may help lessen the disease's impact and, in some cases, even lead to a complete recovery. [3] Support vector machines, decision trees, the k-nearest neighbor algorithm, and so on are just a few of the many algorithms that have been developed and utilized to identify and categorize breast cancer. The majority of studies aimed at detecting and predicting breast cancer employed support vector machines and decision tree algorithms. [17]

1.2 Motivation

According to global statistics, women account for the vast majority of both newly diagnosed cancer cases and cancer-related deaths across the globe. [20] As one of the most common types of cancer, breast cancer is a major health concern worldwide. [1] This article uses four research algorithms to a subset of breast cancer data and analyzes the resulting trends in the development of the health system's big data. This study aims to apply a range of machine-learning approaches for early identification and prevention of breast cancer, the second largest cause of death worldwide among women. [6] Everyone has the right to live in peace and health. Cancer is the deadliest, most debilitating disease that affects humans. Therefore, a definitive method must be discovered to remove these infections and, by extension, protect humans from them. Among the many types of cancer that affect women, breast cancer is one of the most dangerous. [14] For the sake of saving women's lives, it is crucial that early screenings take place often and that a more rapid and accurate diagnosis be made. Primary risk factors for developing breast cancer include female gender, heavy breast tissue, drinking, obesity, and environmental radiation. [27] If caught early, there's a better chance of survival. Machine learning and deep learning may one day be used to identify original breast cancer. [26]

In both developing and developed countries, breast cancer has consistently ranked as the second leading cause of death from any cause. Gene mutation, persistent discomfort, altered stature, altered skin color, and altered skin texture are all telltale signs of breast cancer. Pathologists' ability to make a standardized and final diagnosis is aided by categorization systems, the most frequent of which is a linear one (mild cancer/malignant cancer). Traditional machine learning methods are still frequently used in the breast cancer categorization challenge. There is a great deal of consistency in their grading and diagnosis. [19]

1.The Best Classification System for Predicting Breast Cancer

2.A look of the algorithmic performance of many breast cancer datasets.

3.Choosing the best disease prediction algorithm.

The mechanism that initiates haphazard cell growth, [23] which is at the root of breast cancer, may be inhibited by identifying the faulty cells or other causes of the disease. Male breast cancer, After a delay of many months, [21] masses will be formed by cells that are arranged in the wrong sequence.

1.3 Problem Statement

In this first part, we will focus on breast cancer and its many classifications.

Breast cancer classification is to evaluate potential efficacy or mildness of treatment based on illness classification. Breast cancer categorization relies on nine criteria [18], all of which must be present for a favorable prediction to be made.

1. Generates the layered structures;
2. Determine the accuracy and size of the sample (Cell Size Uniformity);
3. Because cancer cells don't seem to be structurally similar to healthy cells, it's important to estimate cell equality and identify median fluctuations.
4. The cancer cells spread throughout the organ and are intertwined with the normal cells (Marginal Adhesion).
5. Malignant cells are characterized by a lack of regularity and by epithelial cells that have been stretched.
6. Malignant lesions do not have cytoplasm (bare nuclei);
7. In healthy cells, the nucleus always has the same configuration, which is described in detail in point number seven.
8. The nucleolus commonly becomes prominent and highly fragile in normal cells, but in malignancies the chromatin is often coarser (Bland Chromatin). Cancer cells have an abnormally high concentration of nucleoli (Normal Nucleoli);
9. Nine, a quantitative assessment of erupting mitosis. [24] The more favorable the situation, the higher the risk of cancer (Mitoses).

1.4 Research Questions

- Analyzing breast cancer: what's the deal?
- How do we analyze breast cancer?
- How does analyzing breast cancer help?
- How does machine learning fail in its analysis of breast cancer?
- How should the Breast Cancer dataset be trained and preprocessed?
- I was wondering if you could tell me what areas of study this project may expand into.
- When applying several algorithms on the Model, how do they interact with one another?
- How do you describe the machine learning methods that were used to this study?
- Where do difficulties arise while carrying this out?
- How can I get access to all that is available?
- How do the algorithms make use of these variables?
- In this project, how many different algorithms did you use?
- For what reason does the best result come from which algorithm?
- How does this study vary significantly from others?
- What do you want to find as a result of your study?

1.5 Proposed Solution

We want to publish a research article in this area, since this is an exercise in scientific inquiry, and we hope to do so after we have achieved sufficient precision. Numerous researchers have dedicated time and resources to this question, and they have access to high-quality data sets that may help them determine whether or not a certain finding is indicative of breast cancer. To determine whether or not machine learning algorithms can accurately predict or diagnose breast cancer, several have been tried. Our primary objective is to combine all of the available algorithms into a single study to determine which machine learning methods provide the best results. And which method will be the most effective in detecting breast cancer?

Therefore, we can state that the key distinction between our study and that of other researchers is that we are attempting to determine the best accurate machine learning approach for identifying or determining whether any given patient has a possibility of getting breast cancer. Breast cancer may be either malignant or benign, and distinguishing between the two is the primary focus here. When looking for breast cancer, these two factors are crucial. In this study, we want to develop 6 algorithms to compare and select the best algorithm for best machine learning approach that may help us locate or predict breast cancer. Increased precision is what we anticipate. The algorithms that provide the best accuracy for almost the highest accuracy will be those developed using machine learning approaches, which we will use to identify breast cancer. Ultimately, we want to submit our study to a publication or present it at a conference, but that depends on how much more precise our results get.

1.6 Conclusion

In the first chapter of this report, we provided an overview of the whole study project. The associated research that informed this study were covered in Chapter 2. The methods of the study will be detailed in the next chapter. In Chapter 4, we detailed the algorithms we implemented. The outcomes and conclusions of the experiments were described in the previous chapter.

CHAPTER 2

Literature Review

2.1 Introduction

When it comes to cancers that affect women, breast cancer is by far the most common. Much time and effort has been spent studying it. The success of breast cancer therapy has facilitated progress in the study of other types of cancer. Breast cancer treatment has advanced significantly since the disease's discovery. Still, similar research and treatments have been used for quite some time. Breast cancer is a topic that has been discussed for quite some time. To provide only one example, the Edwin Smith Surgical Papyrus [13] includes descriptions of patients with breast cancer. Breasts were often used as votive offerings to the Greek god of healing. This medical book dates back to about 3,000 and 2,500 AD, during which time physicians used surgical incisions to remove malignancies. Hippocrates, writing in the early 400s A.D., describes the progression of breast cancer [22]. As more is learned about breast cancer, clinicians are forced to develop more specific approaches to treatment. Now, experts classify it as a spectrum condition with a wide range of subtypes, each with its own defining characteristics and symptoms. Breast cancer diagnosis and the identification of specific genes involved in the disease mark the beginning of a new line of treatment. Specialized exams may provide additional information regarding breast cancer, even to medical professionals. To verify which genes are active in a tumor subtype, for instance, researchers may use Oncotype DX's gene feature analysis. [25] Doctors will determine which individuals with early breast cancer may be treated without chemotherapy and which will need it.

They hypothesized that stopping menstruation could increase the risk of developing breast cancer. This may have been the starting point for associating cancer with older people. In the early Middle Ages, advances in medicine were linked to emerging religious doctrines. Many Christians came to the conclusion [28] that the practice was cruel and unjust and instead advocated for a secret cure. Meanwhile, Islamic doctors were reading up on breast cancer through the lens of ancient Greek medicine. By the time the Renaissance rolled around, surgery had taken a fresh look at the human body. John Hunter, the Scottish pioneer of exploratory surgery, identified lymph as a risk factor for developing breast cancer. Lymph contains white blood cells [16] that transport and store fluid throughout the body. Surgeons continued to perform lumpectomies without putting patients under. Surgeons aimed to reliably and rapidly improve patient outcomes.

The following are all of the variables that have been linked to an increased chance of developing breast cancer. However, most cases of breast cancer cannot be traced back to a single cause. Consult a medical professional to discuss the specific danger. Rank, Years, Breast cancer incidence increases with a woman's age. [11] Over eighty percent of breast cancers occur in women over the age of fifty. Family history of breast cancer treatment. A woman with a single breast has an increased risk of developing breast cancer in the other breast. The chronic disease of breast cancer. A woman's chance of developing breast cancer increases if her mother, sister, or young daughter has been diagnosed with the disease (before 40). The risk of breast cancer in other family members may also be affected. Inorganic variables. Women with certain genetic abnormalities, most notably changes in the BRCA1 and BRCA2 genes, have a higher lifetime chance of getting breast cancer. [16] The chance of developing breast cancer may also be increased by having certain other gene variants. Menstruation and its historical context in young girls. The risk of breast cancer in older mothers is highest when they are having their first kid. Women who start menstruating at a younger age (until 12 years) postmenopausal women (after 55 years of age) Women who have never given birth face additional challenges..

2.2 Literature Review

After lung and blood cancers, breast cancer is one of the most lethal illnesses worldwide. This horrible condition affects a large number of women and has the potential to inflict significant physical harm. Many scientists, using a wide range of methods and algorithms, are now investigating this issue. Some researchers have attempted to categorize breast cancer using machine learning. They used two distinct classifiers for this purpose: the Naive Bayes (NB) classifier and the k closest neighbor (KNN) classifier. Both are very accurate; the NB algorithm achieves 96.19% while the KNN method achieves 97.51%. [3] We now know that several studies in 2016 employed the Support Vector Machine (SVM), KNN, Decision Tree (C4.5), and NB algorithms to determine breast cancer risk. The SVM algorithm can predict breast cancer risk with 97.13 percent precision. Researchers in another study [2] utilized an automated MRI data set to determine whether breast tumors were present. For this, they used the dataset's photos for a discrete wavelet transform (DWT) feature extraction. They then utilized the SVM technique to look for breast tumors after feature extraction. A precision of 98.03% is achieved by the model. Using the Wisconsin information, another team of researchers classified breast cancer cases to find better methods of treatment [12]. With the use of SVM, KNN, CNN classifiers, Logistic Regression, and Random forest algorithms, they are able to achieve an accuracy of 98% to 99%. Some researchers in 2018 also employed data mining methods to develop a breast cancer survey [7]. To conduct the study, clustering methods were employed. [8] In August of 2020, a team of scientists will analyze the efficacy of several machine learning methods for breast cancer prediction. They used accuracy levels of 98.2% using NB, 98.5% using J48, and 98.8% using KNN. [9] Some academics presented their work on predicting breast cancer risk using XG Boost and Random forest algorithms at a 2020 conference in India. They analyzed 275 cases using 12 characteristics. Using the Random Forest method, they achieved 74.73% accuracy, while XG Boost achieved 73.63%. [10] In order to assess the efficacy of soft classification strategies in the identification of breast cancer, researchers in August 2020 compared the SVM algorithm to five other ML algorithms: KNN, LDA (Linear Discriminant Analysis), NB, LR, and Decision tree classifier (CART). On the other hand, the SVM method yielded the best results. [15] It was reported in 2018 that a researcher has utilized the SVM method in conjunction with machine learning to detect breast cancer, with an accuracy of 97.9 percent. [29]

2.3 Research Summary

The primary objective of this study is to use a breast cancer-related data collection to determine which features of the dataset are necessary for usage with various machine learning algorithms. Models and methods vary depending on the specific machine learning algorithm being used. In addition, we want to test the accuracy with which six distinct machine learning algorithms can detect breast cancer. Investigating whether or not a patient has breast cancer risk factors is our primary objective. The effectiveness of the machine learning algorithms on the dataset will also be assessed. Next, we'll evaluate each method side-by-side to see which machine learning technique yields the most accurate prediction for spotting breast cancer. There have been numerous academics working on specific algorithms, but our goal is to use six distinct algorithms in a single study so that we can better understand and compare them. And what variables affect the outcome of breast cancer forecasting. There are many different kinds of data that have been used by various researchers in their studies. When a patient presents with several symptoms that are also included in the data set, researchers strive to train the data set to predict whether or not the patient has breast cancer based on the number of symptoms that match.

2.4 Conclusion

We confront various obstacles in our study. Locating a dataset that met our needs was a significant hurdle. After that, we need to enhance the data set. Our primary focus is on identifying the datasets that are utilized to make diagnoses and prognoses about breast cancer in patients. While numerous data sets are useful for many types of study, not all of them are specifically useful for predicting or detecting breast cancer. Finding a data collection that can be used as-is and then adjusting it is a necessary step. We gathered this information from Kaggle and from other researchers. Due to the need to make the data set accessible to various machine learning algorithms, we must then process the data and alter certain values in accordance with our study. Since we want to employ six distinct machine learning algorithms in this study, we must ensure that our data is suitable for use by each of them. Once we obtain the data set, we will need to modify it by removing unnecessary columns and values and replacing others. The next step is to choose which of six possible algorithms we will use and how well it integrates with the work of other researchers. The Wisconsin breast cancer dataset is quite well-used. [9][5][4] The references of researchers who have worked with the Wisconsin breast cancer dataset have been put here. This investigation makes use of a substantial quantity Of information.

CHAPTER 3

Research Methodology

3.1 Introduction

Here, we'll talk about the research process as a whole, from start to finish. There is a unique approach to addressing each kind of analysis. The data was first obtained from the Kaggle Wisconsin breast cancer dataset. After that, we do the dataset and process the data, removing the columns and null values that we won't be using. The next step is to decide upon the machine learning algorithm to be used. We've previously said that we'll be using no less than six distinct machine learning algorithms; hence, we'll need to both choose an appropriate method and construct a suitable model before we can begin training. That's the premise upon which feature selection operates. The next step is to create a training set and a testing set from the data. Which is the same thing as a test data set and a training data set. Next, we train the data using the train data set, and then we test the model using a subset of the data. Additionally, the model's correctness may be determined when it is tested. This precision allows us to predict who, if anybody, will be diagnosed with breast cancer. We've used our standard process flowchart to get an overview, but we'll go into some of the algorithms in more detail, including equations and diagrams to help you follow along.

A basic overview of the research process and its associated workflow is provided below.

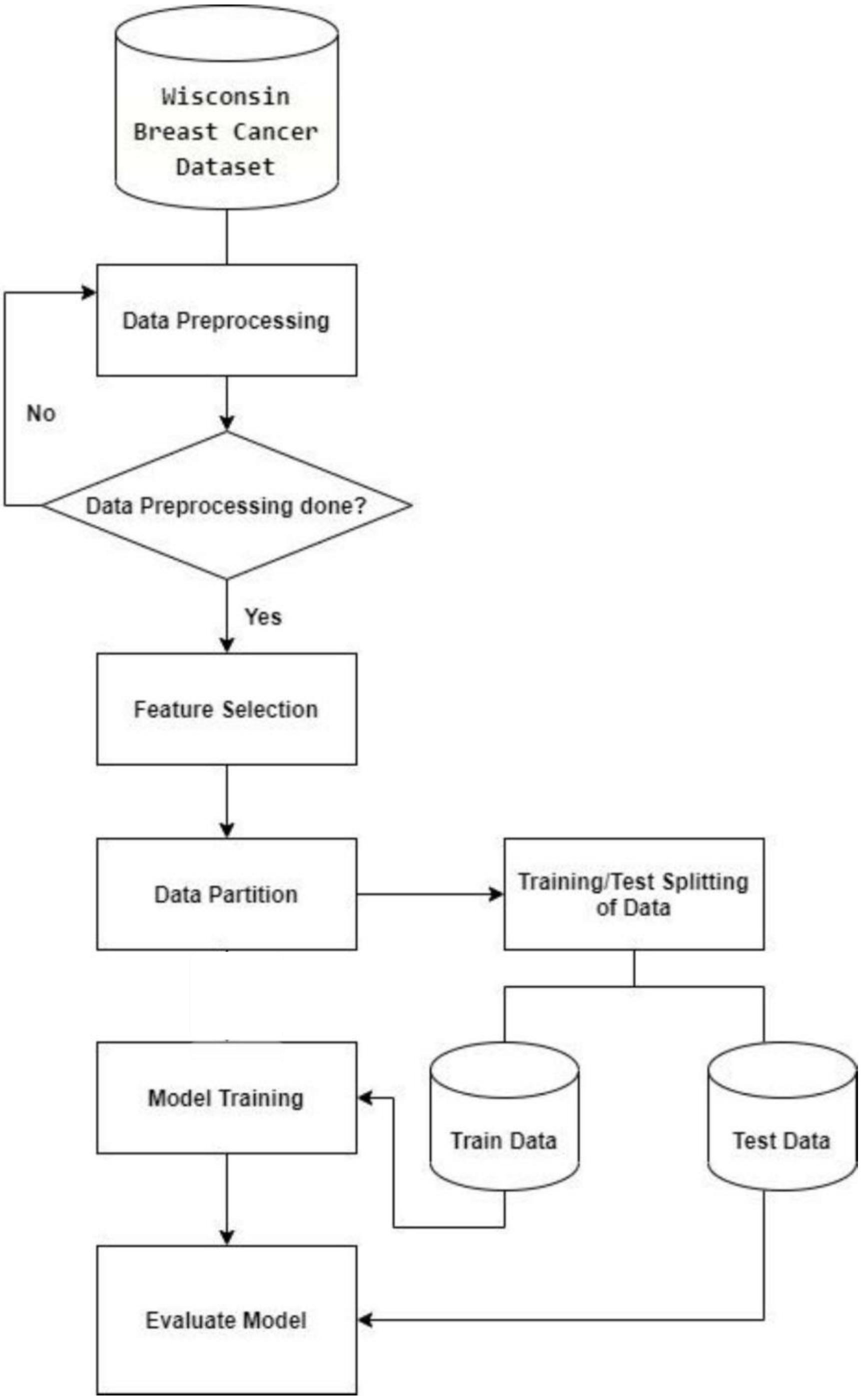


Figure 3.1.1 Workflow for Breast cancer detection

3.2 Research Subject and Instrumentation

First, we've covered the conceptual and theoretical groundwork for finding breast cancer. Computers with high-end graphics processing units (GPUs) and other hardware instruments are required for running machine learning models. The necessary equipment for this model is outlined below.

Hardware and Software:

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

Development Tools:

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

3.3 Data collection and Data preprocessing

We have modified the Kaggle dataset somewhat to meet our needs. The dataset then has to be used for each method. Each of the six algorithms we're using has its own unique methodology, so we need to preprocess the data in the right way before trying to fit it into our model.

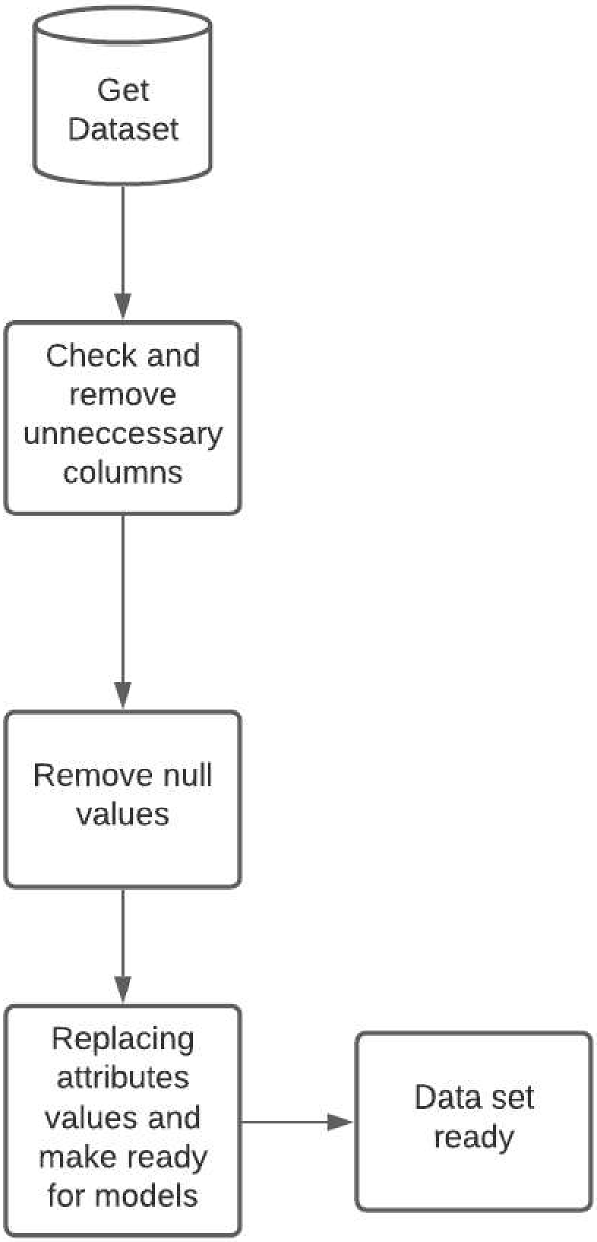

Get Dataset
Check and remove unneccessary columns
Remove null values
Replacing attributes values and make ready for models
Data set ready

Figure 3.3.1: Steps of Dataset Pre-processing

3.3.1 Get Dataset:

We've already said that we collected this information from a wide range of sources, but it's important to note that various researchers choose to focus on different subsets of the overall data pool. Because we want to utilize the same data set with different algorithms to determine whether or not each given patient in the human population has breast cancer, we must choose the data set with great care. For this reason, we need to be selective when selecting the dataset, and we must also preprocess the data appropriately.

3.3.2 Check and remove unnecessary columns:

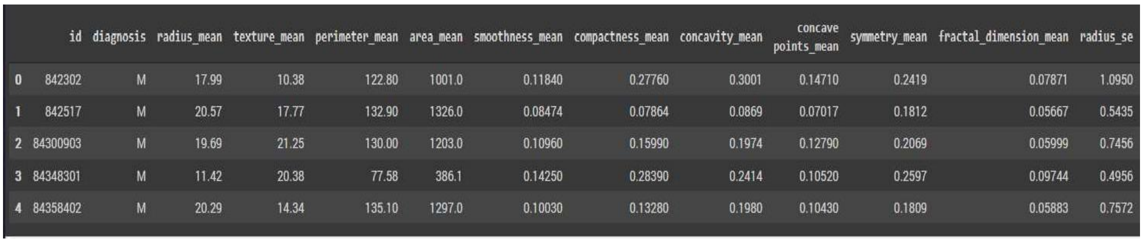
The removal of unnecessary columns is a very important thing for data preprocessing. We have to check if there is any null value or not. The null value actually decreases the accuracy level, so removing the null value will give us a better result because if there is no null value it will help to train the data set properly and give us more accurate result. We also need to eliminate certain columns since our model cannot accommodate them all.

3.3.3 Removing null values and Replacing values:

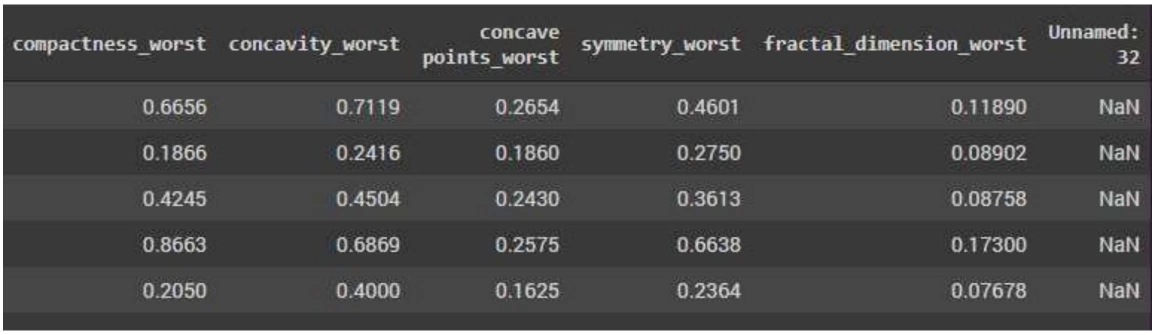
We have previously shown why it's crucial to clean up the data set by deleting any instances of nulls before training with new algorithms: if they're there, the new algorithms won't know how to handle the data and will fail to provide accurate results. And occasionally the algorithm model will provide a less accurate answer for these null values. Due to the possibility of missing data or improper data entry, such instances must be weeded out. When testing the algorithm with this dataset, better accuracy may be seen if all the null values are first removed in a suitable manner, and then training is attacked. Our data collection has a diagnostic property for column names, with two possible values ('M':1 and 'B':0). To fix this, we must switch the M values to 1 and the B values to 0.

3.3.4 Purified data:

Pea processed or purified data is the result of completing all of the phases of data preparation, including as checking for and deleting redundant columns and invalid or missing data. The data after preprocessing will have a more polished appearance. Even though the raw data has a wide variety of elements, after being pre-processed, it takes on a more uniform appearance that makes it simpler to train computers with. Therefore, what we really feed the computer during its training phase is the results of the preprocessing.

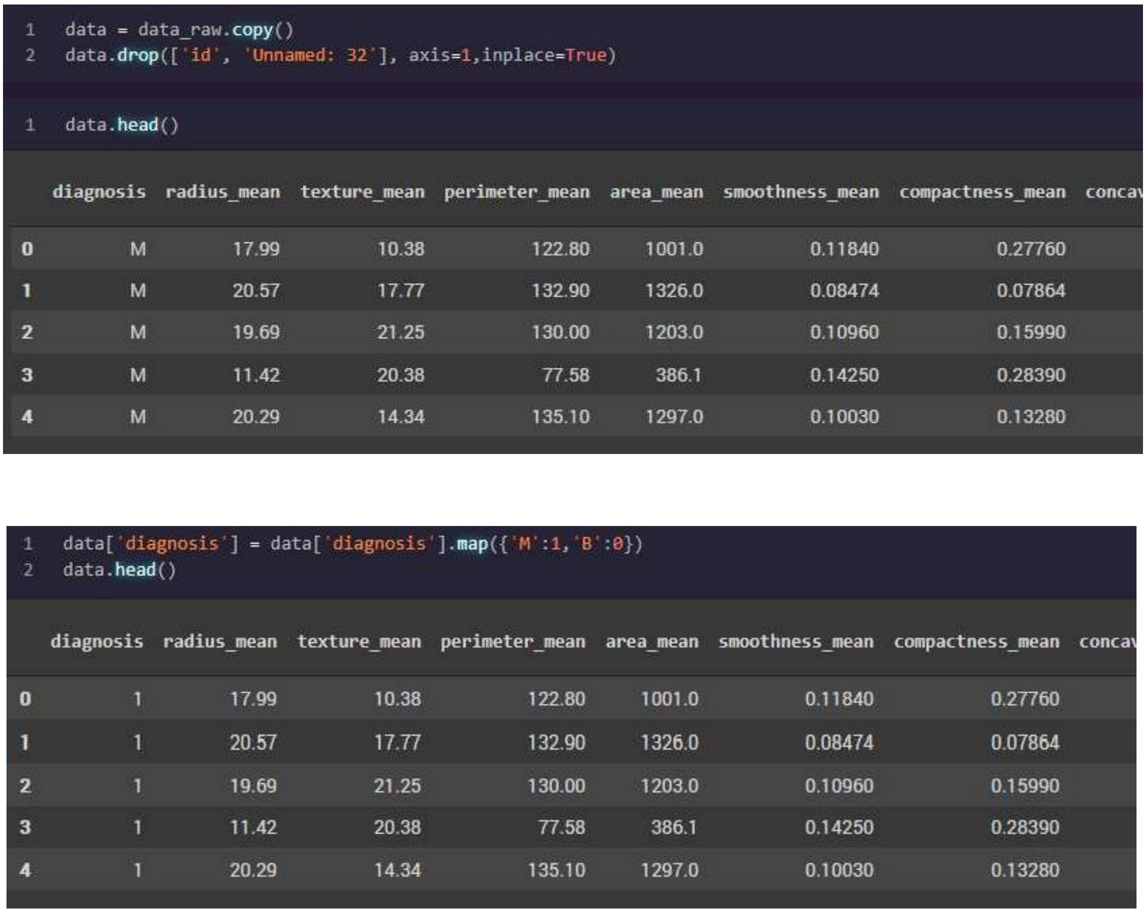


	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal_dimension_mean	radius_se
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419	0.07871	1.0950
1	842517	M	20.57	17.77	132.90	1326.0	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



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

```
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	

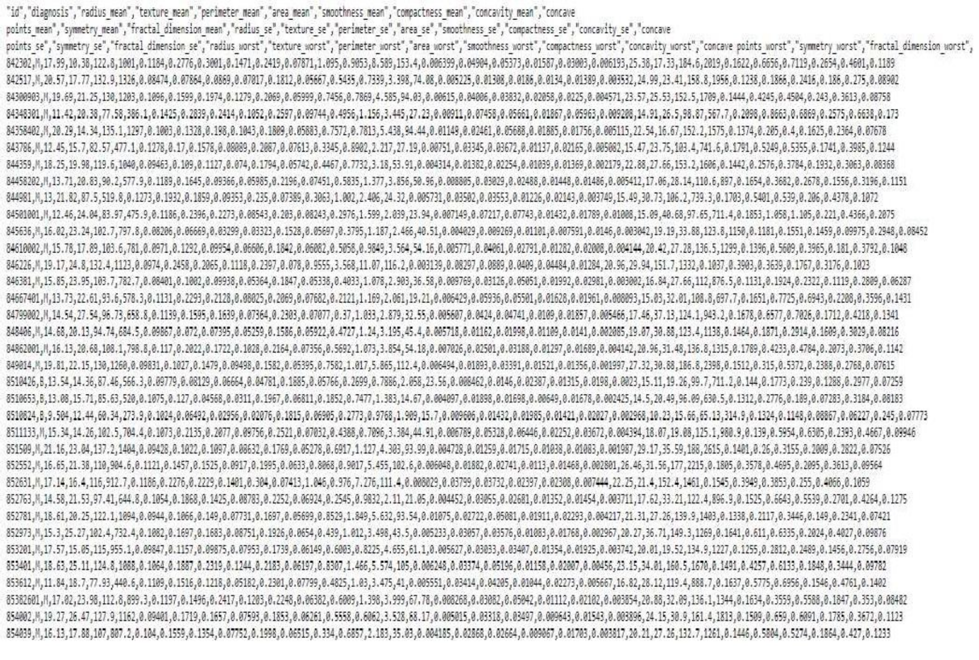
```
data['diagnosis'] = data['diagnosis'].map({'M':1,'B':0})
data.head()
```

	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 massive quantity of data 1. Included in these numbers are 31 subsets labeled things like ['diagnosis', 'radius mean,' 'texture mean,' 'perimeter mean,' 'area mean,''smoothness mean,' 'compactness mean,' 'concavity mean,' 'concave points mean,''symmetry mean,' and so on. 'fractal dimension mean', ''radius,'' ''texture,'' ''perimeter,'' ''area,'' ''smoothness,'' ''compactness,'' ''concavity,'' ''concave points,'' ''convexity,'' ''convex points,'' 'symmetry se', 'fractal dimension se', worst radius, worst texture, worst perimeter, worst area, worst smoothness, worst compactness, worst concavity, worst concave points, worst symmetry, worst fractal dimension]. Here's a quick summary of the information in our database.



```
"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",
842302,M,17.99,10.38,122.8,1001,0.1184,0.2776,0.3001,0.1471,0.2419,0.07871,1.095,0.9053,8.589,153.4,0.006399,0.04904,0.05373,0.01587,0.03003,0.006193,25.38,17.33,184.6,2019,0.1622,0.6656,0.7119,0.2654,0.4601,0.1189
842517,M,20.57,17.77,132.9,1326,0.08474,0.07864,0.0869,0.07017,0.1812,0.05667,0.5435,0.7339,3.398,74.08,0.005225,0.01308,0.0186,0.0134,0.01389,0.003532,24.99,23.41,158.8,1956,0.1238,0.1866,0.2416,0.186,0.275,0.08902
84300903,M,19.69,21.25,130,1203,0.1096,0.1599,0.1974,0.1279,0.2069,0.05999,0.7456,0.7869,4.585,94.03,0.00615,0.04006,0.03832,0.02058,0.0225,0.004571,23.57,25.53,152.5,1709,0.1444,0.4245,0.4504,0.243,0.3613,0.08758
84348301,M,11.42,20.38,77.58,386.1,0.1425,0.2839,0.2414,0.1052,0.2597,0.09744,0.4956,1.156,3.445,27.23,0.00911,0.07458,0.05661,0.01867,0.05963,0.009208,14.91,26.5,98.87,567.7,0.2098,0.8663,0.6869,0.2575,0.6638,0.173
84358402,M,20.29,14.34,135.1,1297,0.1003,0.1328,0.198,0.1043,0.1809,0.05883,0.7572,0.7813,5.438,94.44,0.01149,0.02461,0.05688,0.01885,0.01756,0.005115,22.54,16.67,152.2,1575,0.1374,0.205,0.4,0.1625,0.2364,0.07678
843786,M,12.45,15.7,82.57,477.1,0.1278,0.17,0.1578,0.08089,0.2087,0.07613,0.3345,0.8902,2.217,27.19,0.00751,0.03345,0.03672,0.01137,0.02165,0.005082,15.47,23.75,103.4,741.6,0.1791,0.5249,0.5355,0.1741,0.3985,0.1244
844359,M,18.25,19.98,119.6,1040,0.09463,0.109,0.1127,0.074,0.1794,0.05742,0.4467,0.7732,3.18,53.91,0.004314,0.01382,0.02254,0.01039,0.01369,0.002179,22.88,27.66,153.2,1606,0.1442,0.2576,0.3784,0.1932,0.3063,0.08368
84458202,M,13.71,20.83,90.2,577.9,0.1189,0.1645,0.09366,0.05985,0.2196,0.07451,0.5835,1.377,3.856,50.96,0.008805,0.03029,0.02488,0.01448,0.01486,0.005412,17.06,28.14,110.6,897,0.1654,0.3682,0.2678,0.1556,0.3196,0.1151
844981,M,13,21.82,87.5,519.8,0.1273,0.1932,0.1859,0.09353,0.235,0.07389,0.3063,1.002,2.406,24.32,0.005731,0.03502,0.03553,0.01226,0.02143,0.003749,15.49,30.73,106.2,739.3,0.1703,0.5401,0.539,0.206,0.4378,0.1072
84501001,M,12.46,24.04,83.97,475.9,0.1186,0.2396,0.2273,0.08543,0.203,0.08243,0.2976,1.599,2.039,23.94,0.007149,0.07217,0.07743,0.01432,0.01789,0.01008,15.09,40.68,97.65,711.4,0.1853,1.058,1.105,0.221,0.4366,0.2075
845636,M,16.02,23.24,102.7,797.8,0.08206,0.06669,0.03299,0.03323,0.1528,0.05697,0.3795,1.187,2.466,40.51,0.004029,0.009269,0.01101,0.007591,0.0146,0.003042,19.19,33.88,123.8,1150,0.1181,0.1551,0.1459,0.09975,0.2948,0.08452
84610002,M,15.78,17.89,103.6,781,0.0971,0.1292,0.09954,0.06606,0.1842,0.06082,0.5058,0.9849,3.564,54.16,0.005771,0.04061,0.02791,0.01282,0.02008,0.004144,20.42,27.28,136.5,1299,0.1396,0.5609,0.3965,0.181,0.3792,0.1048
846226,M,19.17,24.8,132.4,1123,0.0974,0.2458,0.2065,0.1118,0.2397,0.078,0.9555,3.568,11.07,116.2,0.003139,0.08297,0.0889,0.0409,0.04484,0.01284,20.96,29.94,151.7,1332,0.1037,0.3903,0.3639,0.1767,0.3176,0.1023
846381,M,15.85,23.95,103.7,782.7,0.08401,0.1002,0.09938,0.05364,0.1847,0.05338,0.4033,1.078,2.903,36.58,0.009769,0.03126,0.05051,0.01992,0.02981,0.003002,16.84,27.66,112,876.5,0.1131,0.1924,0.2322,0.1119,0.2809,0.06287
84667401,M,13.73,22.61,93.6,578.3,0.1131,0.2293,0.2128,0.08025,0.2069,0.07682,0.2121,1.169,2.061,19.21,0.006429,0.05936,0.05501,0.01628,0.01961,0.008093,15.03,32.01,108.8,697.7,0.1651,0.7725,0.6943,0.2208,0.3596,0.1431
84799002,M,14.54,27.54,96.73,658.8,0.1139,0.1595,0.1639,0.07364,0.2303,0.07077,0.37,1.033,2.879,32.55,0.005607,0.0424,0.04741,0.0109,0.01857,0.005466,17.46,37.13,124.1,943.2,0.1678,0.6577,0.7026,0.1712,0.4218,0.1341
848406,M,14.68,20.13,94.74,684.5,0.09867,0.072,0.07395,0.05259,0.1586,0.05922,0.4727,1.24,3.195,45.4,0.005718,0.01162,0.01998,0.01109,0.0141,0.002085,19.07,30.88,123.4,1138,0.1464,0.1871,0.2914,0.1609,0.3029,0.08216
84862001,M,16.13,20.68,108.1,798.8,0.117,0.2022,0.1722,0.1028,0.2164,0.07356,0.5692,1.073,3.854,54.18,0.007026,0.02501,0.03188,0.01297,0.01689,0.004142,20.96,31.48,136.8,1315,0.1789,0.4233,0.4784,0.2073,0.3706,0.1142
849014,M,19.81,22.15,130,1260,0.09831,0.1027,0.1479,0.09498,0.1582,0.05395,0.7582,1.017,5.865,112.4,0.006494,0.01893,0.03391,0.01521,0.01356,0.001997,27.32,30.88,186.8,2398,0.1512,0.315,0.5372,0.2388,0.2768,0.07615
8510426,B,13.54,14.36,87.46,566.3,0.09779,0.08129,0.06664,0.04781,0.1885,0.05766,0.2699,0.7886,2.058,23.56,0.008462,0.0146,0.02387,0.01315,0.0198,0.0023,15.11,19.26,99.7,711.2,0.144,0.1773,0.239,0.1288,0.2977,0.07259
8510653,B,13.08,15.71,85.63,520,0.1075,0.127,0.04568,0.0311,0.1967,0.06811,0.1852,0.7477,1.383,14.67,0.004097,0.01898,0.01698,0.00649,0.01678,0.002425,14.5,20.49,96.09,630.5,0.1312,0.2776,0.189,0.07283,0.3184,0.08183
8510824,B,9.504,12.44,60.34,273.9,0.1024,0.06492,0.02956,0.02076,0.1815,0.06905,0.2773,0.9768,1.909,15.7,0.009606,0.01432,0.01985,0.01421,0.02027,0.002968,10.23,15.66,65.13,314.9,0.1324,0.1148,0.08867,0.06227,0.245,0.07773
8511133,M,15.34,14.26,102.5,704.4,0.1073,0.2135,0.2077,0.09756,0.2521,0.07032,0.4388,0.7096,3.384,44.91,0.006789,0.05328,0.06446,0.02252,0.03672,0.004394,18.07,19.08,125.1,980.9,0.139,0.5954,0.6305,0.2393,0.4667,0.09946
851509,M,21.16,23.04,137.2,1404,0.09428,0.1022,0.1097,0.08632,0.1769,0.05278,0.6917,1.127,4.303,93.99,0.004728,0.01259,0.01715,0.01038,0.01083,0.001987,29.17,35.59,188,2615,0.1401,0.26,0.3155,0.2009,0.2822,0.07526
852552,M,16.65,21.38,110,904.6,0.1121,0.1457,0.1525,0.0917,0.1995,0.0633,0.8068,0.9017,5.455,102.6,0.006048,0.01882,0.02741,0.0113,0.01468,0.002801,26.46,31.56,177,2215,0.1805,0.3578,0.4695,0.2095,0.3613,0.09564
852631,M,17.14,16.4,116,912.7,0.1186,0.2276,0.2229,0.1401,0.304,0.07413,1.046,0.976,7.276,111.4,0.008029,0.03799,0.03732,0.02397,0.02308,0.007444,22.25,21.4,152.4,1461,0.1545,0.3949,0.3853,0.255,0.4066,0.1059
852763,M,14.58,21.53,97.41,644.8,0.1054,0.1868,0.1425,0.08783,0.2252,0.06924,0.2545,0.9832,2.11,21.05,0.004452,0.03055,0.02681,0.01352,0.01454,0.003711,17.62,33.21,122.4,896.9,0.1525,0.6643,0.5539,0.2701,0.4264,0.1275
852781,M,18.61,20.25,122.1,1094,0.0944,0.1066,0.149,0.07731,0.1697,0.05699,0.8529,1.849,5.632,93.54,0.01075,0.02722,0.05081,0.01911,0.02293,0.004217,21.31,27.26,139.9,1403,0.1338,0.2117,0.3446,0.149,0.2341,0.07421
852973,M,15.3,25.27,102.4,732.4,0.1082,0.1697,0.1683,0.08751,0.1926,0.0654,0.439,1.012,3.498,43.5,0.005233,0.03057,0.03576,0.01083,0.01768,0.002967,20.27,36.71,149.3,1269,0.1641,0.611,0.6335,0.2024,0.4027,0.09876
853201,M,17.57,15.05,115,955.1,0.09847,0.1157,0.09875,0.07953,0.1739,0.06149,0.6003,0.8225,4.655,61.1,0.005627,0.03033,0.03407,0.01354,0.01925,0.003742,20.01,19.52,134.9,1227,0.1255,0.2812,0.2489,0.1456,0.2756,0.07919
853401,M,18.63,25.11,124.8,1088,0.1064,0.1887,0.2319,0.1244,0.2183,0.06197,0.8307,1.466,5.574,105,0.006248,0.03374,0.05196,0.01158,0.02007,0.00456,23.15,34.01,160.5,1670,0.1491,0.4257,0.6133,0.1848,0.3444,0.09782
853612,M,11.84,18.7,77.93,440.6,0.1109,0.1516,0.1218,0.05182,0.2301,0.07799,0.4825,1.03,3.475,41,0.005551,0.03414,0.04205,0.01044,0.02273,0.005667,16.82,28.12,119.4,888.7,0.1637,0.5775,0.6956,0.1546,0.4761,0.1402
85382601,M,17.02,23.98,112.8,899.3,0.1197,0.1496,0.2417,0.1203,0.2248,0.06382,0.6009,1.398,3.999,67.78,0.008268,0.03082,0.05042,0.01112,0.02102,0.003854,20.88,32.09,136.1,1344,0.1634,0.3559,0.5588,0.1847,0.353,0.08482
854002,M,19.27,26.47,127.9,1162,0.09401,0.1719,0.1657,0.07593,0.1853,0.06261,0.5558,0.6062,3.528,68.17,0.005015,0.03318,0.03497,0.009643,0.01543,0.003896,24.15,30.9,161.4,1813,0.1509,0.659,0.6091,0.1785,0.3672,0.1123
854039,M,16.13,17.88,107,807.2,0.104,0.1559,0.1354,0.07752,0.1998,0.06515,0.334,0.6857,2.183,35.03,0.004185,0.02868,0.02664,0.009067,0.01703,0.003817,20.21,27.26,132.7,1261,0.1446,0.5804,0.5274,0.1864,0.427,0.1233
```

Fig 3.4.1: Dataset Preview

1. First, there are a total of 31 columns.
2. There are around 570 rows in all.
3. Third, we trained our model using 80% of the available data.
4. Fourth, test data makes up 20% of the total.
5. The dataset is a.csv file, which can be opened in Excel.

3.5 Conclusion

We'll cover the whole study process here. Each analysis is handled differently. Kaggle Wisconsin breast cancer dataset provided the data. We next analyze the dataset and remove columns and null values we won't use. Choose a machine learning algorithm next. We'll use six machine learning algorithms, so we'll need to pick a technique and build a model before training. That's feature selection's basis. Next, generate training and testing sets from the data. The same as a test and training data set. Next, we train the data using the train data set and test the model with a subset. Testing can verify the model's accuracy. This accuracy lets us anticipate breast cancer cases. We'll explain several algorithms using equations and diagrams after using our typical process flowchart to obtain an overview.

CHAPTER 4

Theoretical Model

4.1 Introduction

Support Vector Machine (SVM), Random Forest Classifier (RFC), K-Nearest Neighbor (KNN), Gaussian Naive Bayes (GNB), Logistic Regression (LR), and Decision Tree (DT) are just some of the ML algorithms that will be discussed in this chapter, along with their respective implementation procedures in this study.

4.2 SVM (Support Vector Machine)

SVM may be used as a discriminating classifier shown on a hyperplane of separation. SVM is often a well-known regulated master learning law. It is utilized in categorization difficulties regularly. Each element in an n-dimensional space may be represented in the SVM model as a degree that represents how much that space is valued at that given point (where n is many features). If you were to use Associate to explain the difference between the two types of things, you would have no trouble understanding the example provided. This estimate is useful for deducing how the SVM classifier works.

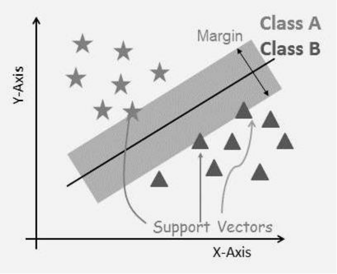


Figure 4.1.1: SVM figure [14]

The primary objective is to isolate data collection as much as possible. The function is the separation of some of the most closely spaced data points. The goal is to calculate the largest possible separation of a hyperplane from a set of vectors that provide support for the plane. Specifically, SVM follows these procedures to locate the nominal maximum hyperplane:

Create hyperplanes to further define the categories. The diagram on the left shows three hyperplanes, one each in yellow, blue, and orange. Black correctly differentiates between the two groups, however the blue and orange are incorrectly placed in the upper division.

Select the appropriate hyperplane from the set of closest distances.

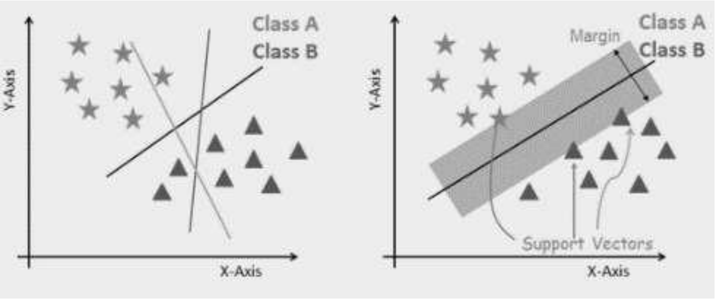


Figure 4.1.2: How SVM works [11]

While Maintaining control over a plane that is both nonlinear and difficult to discern;

As can be shown later, there are potentially insurmountable difficulties associated with linear hyperplanes (left-hand side).

As can be seen in the illustration to the right, SVM employs a kernel technique in this situation to map the original establishment space onto a higher dimensional feature space. The data points are shown along the x-axis and the z-axis, respectively (Z is the squared number of x and y: z=x2=y2). You may now easily distinguish these locations using linear segregation.

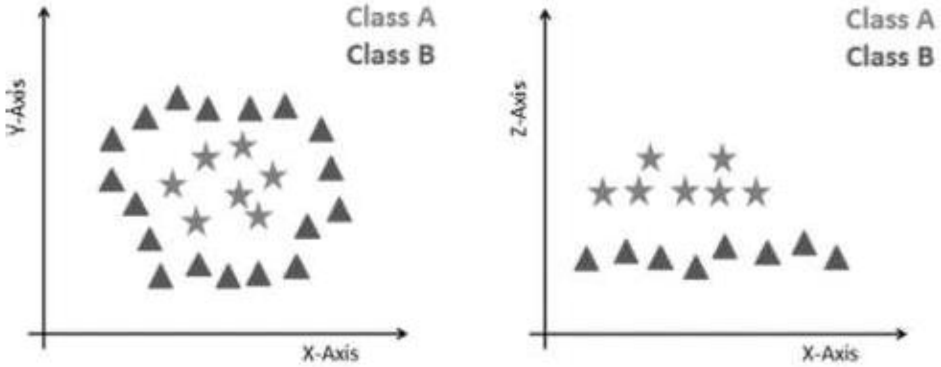


Figure 4.1.3: Managing nonlinear and indistinguishable planes

The SVM algorithm is run in the context of a kernel. The input data space may be converted into the proper format by using a kernel. When it comes to the SVM, a kernel technique is used. In this scenario, the kernel expands a constrained input space. If we add additional dimensions to an issue, we split it up into smaller, more manageable pieces, right? It is best used to separation issues that do not follow a linear pattern. By using the kernel method, you can make a more nuanced sorting.

The dot product of any two operations is a valid linear kernel. Multiplication of the sums inside each pair of input data produces the product between the two vectors. These equations illustrate the mathematical representation:

K (x, xi) = sum (x * xi)

The linear kernel structure of the polynomial would be the most common. Curved or nonlinear input space may be distinguished by the kernel polynomial

.**K (x, xi) = 1 + sum (x * xi) ^d**

In where d represents the degree of polynomial. As a realistic implementation, d=1 comes quite near to that ideal. The degree of difficulty of the learning algorithm is determined by hand.

One of the most well-known kernel functions for use in classification using vector machines is the radial functional kernel. RBF can remap input space even in infinite dimensions.

K (x, xi) = exp (-gamma * sum ((x –xi^2))

Here, gamma is a continuous function between 0 and 1. If the gamma is set too high, it will adapt too much to the training data. A decent anticipated value is often around gamma=0.1. The gamma value must be specified explicitly throughout the whole learning method. The SVM categories provide acceptable accuracy and produce quicker prediction compared to the Naive Bayes method. They tend to rely on a more limited collection of information while making decisions and storing memories. SVM works best when there is a large range and a large margin of separation.

4.3 Random Forest Classifier

It is true that random forests are a supervised analysis algorithm. as well as being applicable to regression analysis. It's possible that no other algorithm is as flexible or user-friendly as this one. A forest is a dense area of trees. A forest's quality increases as its tree population grows. Decision trees are randomly sampled, accuracy is obtained from each tree, and the best answer is selected by a voting process in a random forest. It's also a great measure of the function's worth. Recommendation systems, picture recognition, and feature aggregation are just a few of the many areas where random forests have found success. It can sort reliable creditors, uncover fraud, and even predict illnesses. Choosing pivotal features of a dataset is the foundation of the Boruta method.

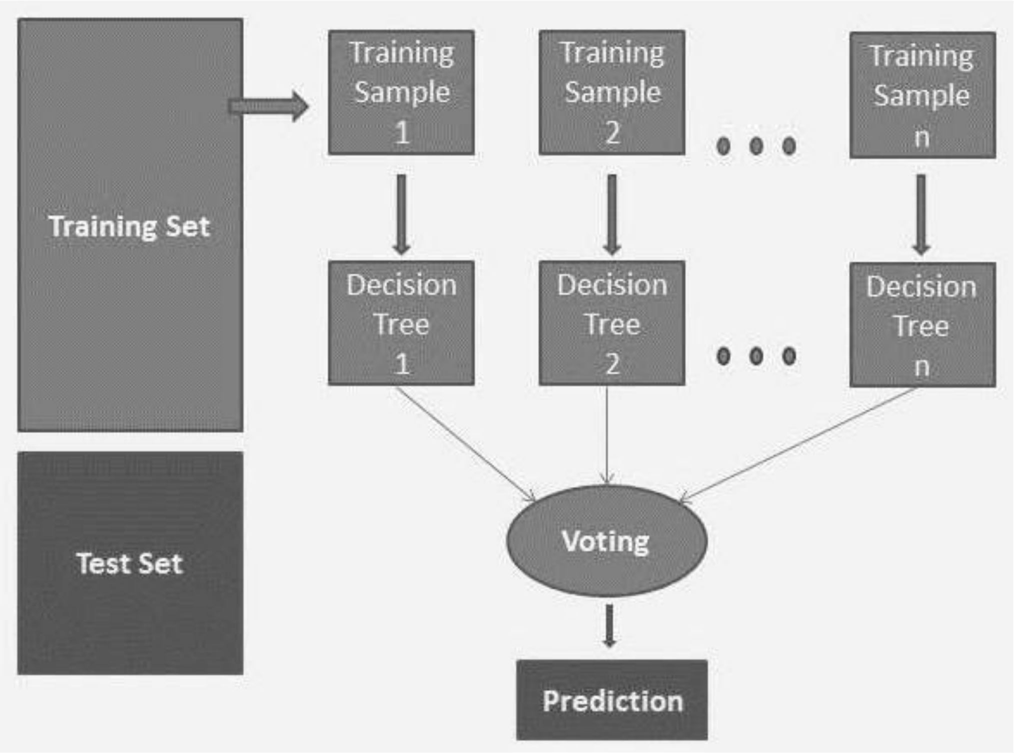


Figure 4.2.1: Steps of how Random forest classifier works

1. The first step is to choose a subset of the data set at random.
2. For each sample, step two is to generate a decision tree and then use that tree to create a prediction. Put each potential conclusion to a vote.
3. Third, make the most popular projection the final one.

Due to the large number of decision trees used in the process, random forests are regarded as a very robust and accurate method. There is no overproduction problem. The most important reason is that every forecast averages out the predictions. Both methods may be used to the classification and regression queries. Similarly, random forest may be used to control outliers. Median values are utilized to supersede the continuous variables, and the average weighted closeness of missing values is calculated. The classifier's parameters may be set with the assistance of the function's relative value, which can be obtained. Using the suggested random forest method, we have used the methods of bootstrapping for tree samples. When a dataset X=x1...x n is used for training with answers Y=y1, y n boosting selects a random sample and substitutes a tree representing the samples for the training set. After everything is processed, we may project onto x' an unknown sample by averaging the projections onto every node in the tree:

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

4.4 K Nearest Neighbor

The K-Nearest Neighbor (KNN) Classifier is one of the simplest algorithms, and yet it is highly effective, flexible, and scalable. KNN has applications in many different disciplines, including economics, medicine, politics, hand-writing recognition, visual recognition, and video recognition.

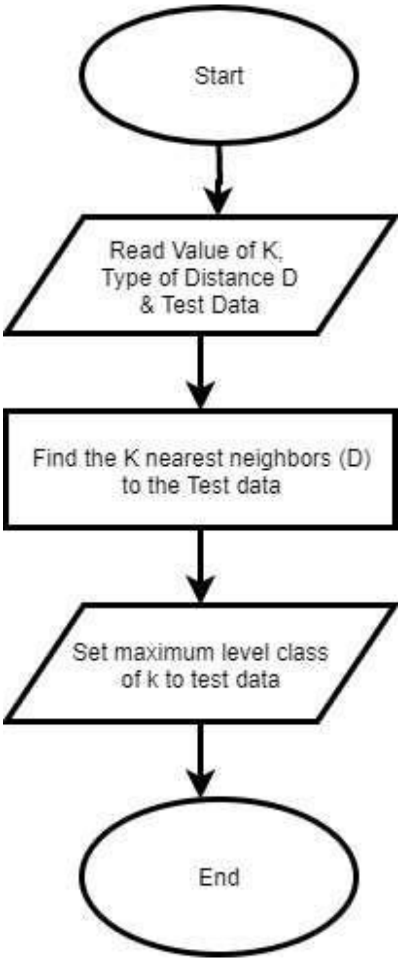


Figure 4.3.1: KNN flowchart

Euclidean metric equation:

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

Probability:

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

Credit ratings allow financial institutions to foresee customers' creditworthiness. Financial firms do a risk assessment prior to extending credit to determine how safe or risky the loan is. In politics, there are two types of potential candidates: those who vote and those who do not. In both regression and classification problems, the KNN technique is utilized. In KNN, the method of comparing features was the main emphasis. The method for learning does not take into account parameters. Means that aren't calculated using a parametric distribution assumption are ignored. Dataset served as a description of the model's setup. Even though most real-time datasets don't adhere to statistical science concepts, this is still incredibly helpful. The algorithm's lack of intelligence guarantees that no specialized data centers for model training will be needed to generate the model. All information gathered for the purpose of preparing for the examination. That cuts down on the time and money needed for training and testing. Remembering and checking one's memory takes time and money. Searching all data points and all information storage locations would take more time in the worst case scenario. In the KNN algorithm, K represents the number of nearest neighbors. The number of nearby neighbors is the most important aspect. When the groupings of classes are 2, K is often an unusual digit. The method is known as the nearest algorithm when K is equal to 1. In this case, everything could not be easier. Let's say mark is expecting a score of P1. The first step was to locate P1 and the label next to it.

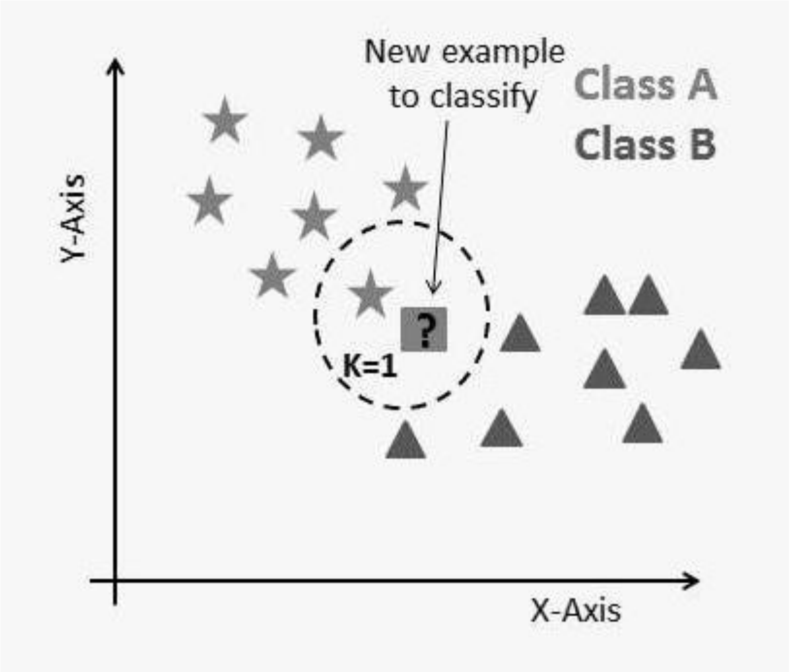


Figure 4.3.2: The process of KNN algorithm

Mark will be making a projection onto point P1. The next step is to determine which of P1's neighbors are in the closest k, and then use that number to determine which category P1 belongs to. The clause is applied to the most popular category and the item's own voting category. The distance between points is determined by methods of distance measures like Euclidean, Hamming and Manhattan as well as Minkowski as the closest points are close. KNN uses the following straightforward procedures:Range estimation

a) Find the nearest vicinity

b) Vote for markings

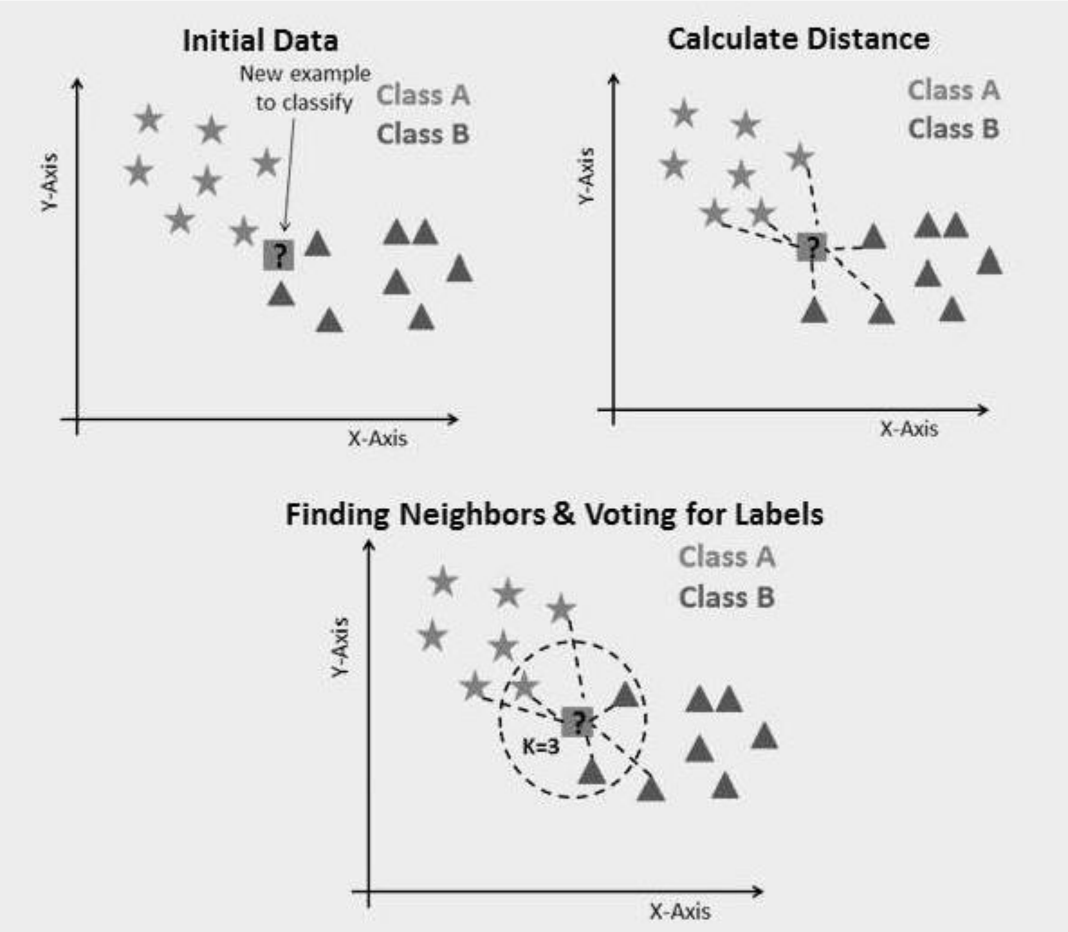


Figure 4.3.3: The process of KNN algorithm

The variable K in the projection model is used to control the projection. In general, you can't find "ideal neighbors" for your data. each dataset has its unique set of requirements. However, with a high enough number of neighbors, the effects of the vibration will become too great to be handled computationally efficiently. Evidence suggests that the most adaptable match occurs when a limited number of nodes are used in the judgment limit, whereas a large number of neighbors indicates less variance and more bias. Whenever the number of participants evens out, data scientists often get an odd number.

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

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

4.5 Gaussian Naïve Bayes

To rate a large range of data, Naive Bayes is the best and simplest method. Naive Bayes classification is useful in a variety of contexts, such as spam filtering, document categorization, sentiment analysis, and recommendation engines. In order to forecast classes that are not explicitly given, it uses Bayes' probability theorem.

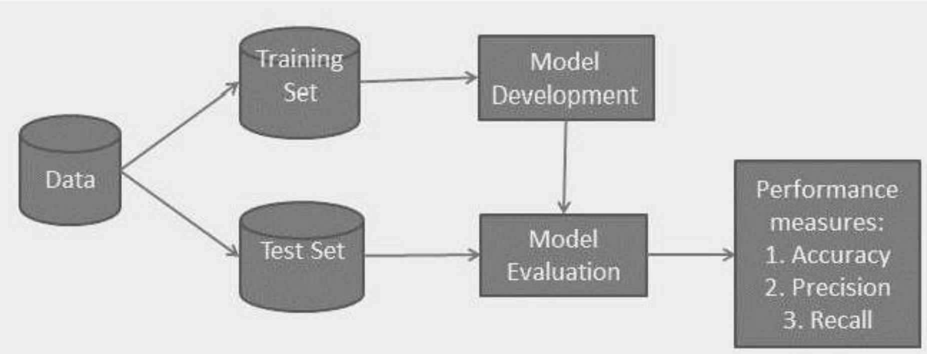


Figure 4.4.1: Naïve bayes working process

The procedure is broken down into two phases: the research phase and the evaluation phase. In the learning stage, the classifier uses the input data to fine-tune its model; in the evaluation stage, the model's performance is evaluated. Results were assessed using many criteria, such as percentage of correct answers, number of errors, frequency of reminders, and overall accuracy. Naive Bayes is a statistical approach for classifying data that is based on the Bayes theorem. One of the simplest supervised algorithms is this one. The Naive Bayes algorithm is the fastest, most accurate, and most trustworthy option. Naive Bayes classifiers can be fast and accurate, even when applied to massive data sets..

Phase 1: Rough Estimates and Some Grades

Phase 2: Establish the Likelihood of Each Class Attribute

Phase 3: Use the Bayes Formula to plug in this value and get the probability.

Phase 4: Determine which group is most likely, given that the response falls into that group

Naive Bayes's approach to classification assumes that a function's impact inside a class is not reliant on any other attributes. For instance, a borrower's salary, work history, age, and location of the credit and transaction all play a role in determining whether or not they're a good candidate for Despite the fact that there is a close relationship between these characteristics, they are examined independently. This naïve assumption simplifies programming yet has the opposite effect. The term "class emancipation" is used to describe this kind of assumption.

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

P(h): the potential for H-theory (regardless of the data). The term for this is "h's prior possibility."

P(D): chance analysis (regardless of the hypothesis). The term "prior opportunity" describes this situation.

P (h |D): Given the D data, a high-probability h hypothesis can be made. A posterior probability would describe this.

P (D| h): possible results from collecting data d, given that H is correct. A posterior probability would describe this.

4.6 Logistics Regression:

Linear Regression (LR) is a mathematical method for predicting two-class binary data. The outcome or criterion variable is essentially binary. As a result, there are only two potential demographics. Depending on the context, TRUE might mean 1 or FALSE, 0. Logistics regression resembles linear regression in many respects. In the language of linear models, we may express that. Linear Regression Equation.

$$y = \beta 0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + +\beta_n X_n$$

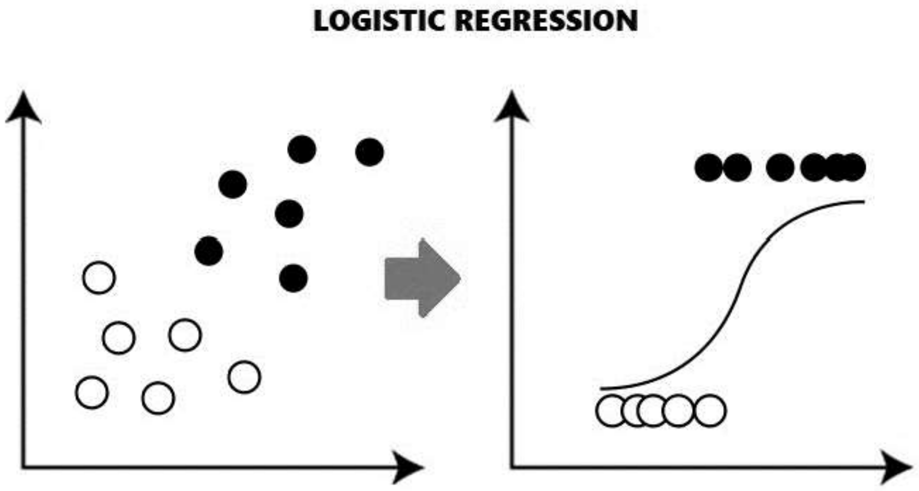


Figure 4.5.1: Logistic Regression

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

Final equation after applying sigmoid function:

$$p = \frac{1}{1 + e^{-(0+\beta_1 X_1+\beta_2 X_2+\cdots++\beta_n X_n)}}$$

4.7 Decision Tree

A decision tree is a kind of flux-chart tree structure, with the core node representing some kind of method (or attribute), the branch representing some kind of decision rule, and the leaf nodes representing the results of applying that law. In a decision tree, the very first node is known as the root node. It depends on the attribute's actual value. A tree is partitioned in a recursive manner into periodic subtrees. You may make selections using this illustrative flowchart format. It's a visual representation, like a diagram, that mimics the way people think. For this reason, decision trees are intuitive.

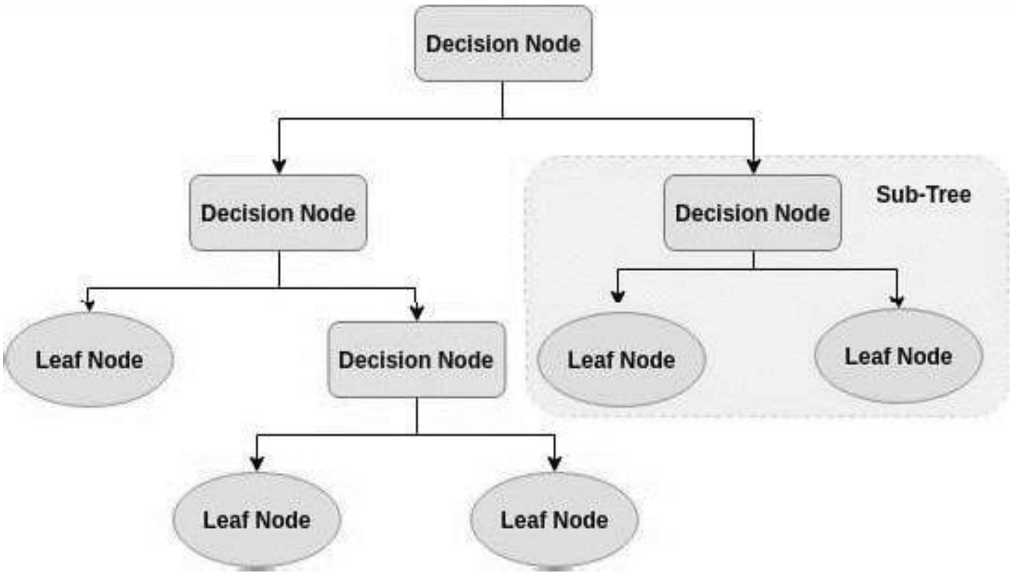


Figure 4.6.1: Decision tree

In the realm of machine learning algorithms, Decision Tree is the white box. Decision-making is built in, and there is no black box in algorithms like neural networks. It requires less time to train than the neural network approach. Time complexity of the decision tree has a major effect on the number of records and the number of characteristics in the data presented. The decision tree does not depend on prior probabilistic beliefs while making its determinations. The accuracy of decision trees is not degraded while working with high-dimensional data. When sorting documents with the ASM, choose the characteristic that works best for you. Judgment node attributes and threaded data collecting. Commences tree creation by iteratively applying the following procedure to each child node until one of the following conditions holds true: Both tuples have the same worth as characteristics. We can no longer identify any remaining features. The violence has stopped.

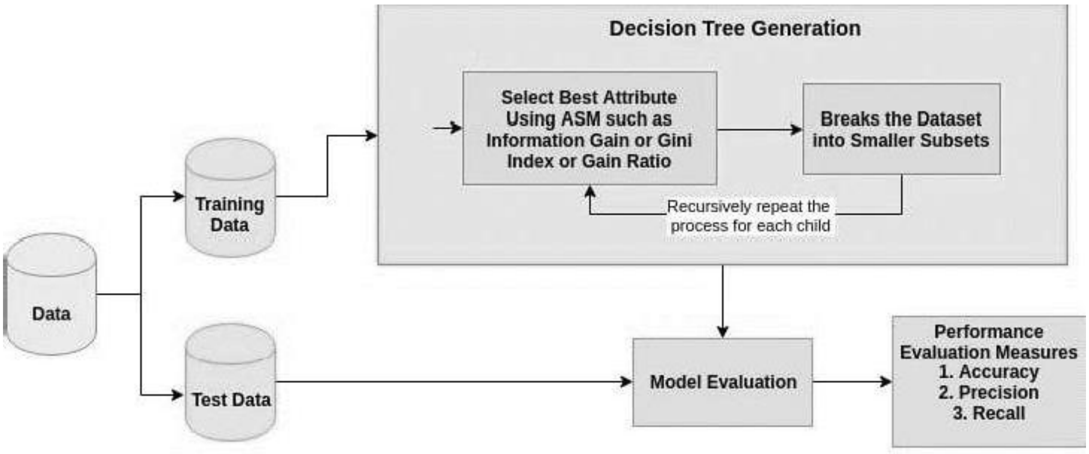

Decision Tree Generation
Select Best Attribute Using ASM such as Information Gain or Gini Index or Gain Ratio
Breaks the Dataset into Smaller Subsets
Recursively repeat the process for each child
Training Data
Data
Test Data
Model Evaluation
Performance Evaluation Measures
1. Accuracy
2. Precision
3. Recall

Figure 4.6.2: Working process of Decision Tree

Attribute Selection Measure:

The measure of characteristics selection is a heuristic for selecting appropriate criteria for partitioning. Also, the criteria for determining where to divide tuples at a given node are well-known. ASM provides a ranking for each characteristic by explaining the provided dataset (s). The top rating is used as a criteria for differentiation. Establishing branching points is also a requirement of the "in progress" characteristic. The most often used performance behaviors are the knowledge benefit, gain ratio, and Gini index.Information Gain:

It's entropy's input impurity word that's evolved. In physics and mathematics, entropy is used to represent unpredictability or impurity in a system. This defect is analogous to the presence of impurity in a set of information theoretic illustrations. The loss of entropy represents a gain in knowledge. After dividing the dataset according to the values of the attribute, the data benefit is the difference between the entropy and the average entropy. Knowledge is used as an advantage in the decision tree algorithm ID3 (Iterative Dichotomies).

$$Info(D) = \Sigma_{i=1}^{m} P^{i}\, log_2\, pi$$

$$Info_A(D) = \sum_{j=1}^{V} \frac{|D_j|}{|D|} X\, Info(D_j)$$

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

$$\Delta\, \mathrm{Gini(A)} \;=\; \mathrm{Gini(D)} - Gini_a(D)$$

4.8 Conclusion

In this section, we discussed the methods by which we implemented ML algorithms such as the Support Vector Machine (SVM), Random Forest Classifier (RFC), K-Nearest Neighbor (KNN), Gaussian Naive Bayes (GNB), Logistic Regression (LR), and Decision Tree (DT).

CHAPTER 5
Experimental Results, Discussion and Conclusion

5.1 Introduction

Evolved is the impure input term for entropy. Entropy is a physical and mathematical concept used to express disorder and impureness. This flaw is like an impurity in an information-theoretic collection of examples. Knowledge is increased via the reduction of entropy. The data benefit is the difference between the entropy and the average entropy after the dataset is divided according to the values of the attribute. ID3 is a decision tree algorithm that takes use of prior knowledge (Iterative Dichotomies).

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

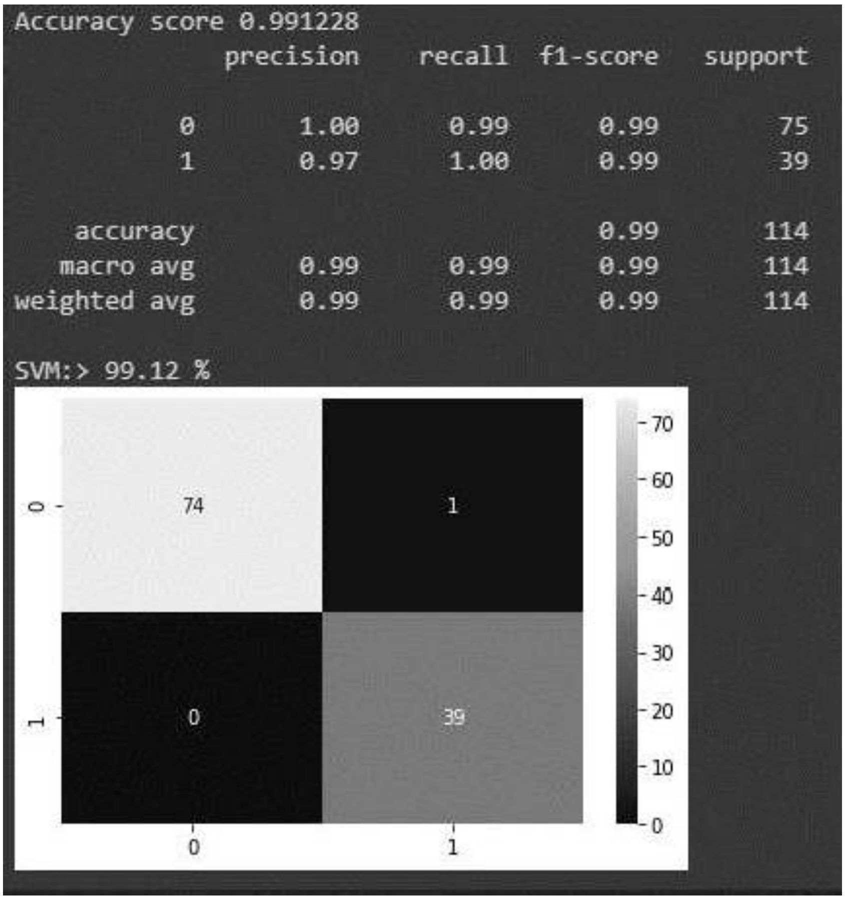
Step-8: Compare which algorithm gives the best result

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

5.2 Experimental Results

We Everybody here is aware that no machine can ever provide a perfect efficiency rate. In a similar vein, we may fine-tune our model's training by adjusting various parameters. However, the results we get from various algorithms are rather satisfactory in terms of accuracy.

We've included some photographs below that offer you a quick look at our study. Precision, recall, f1 score, support, accuracy, and heatmap are all shown in these images.



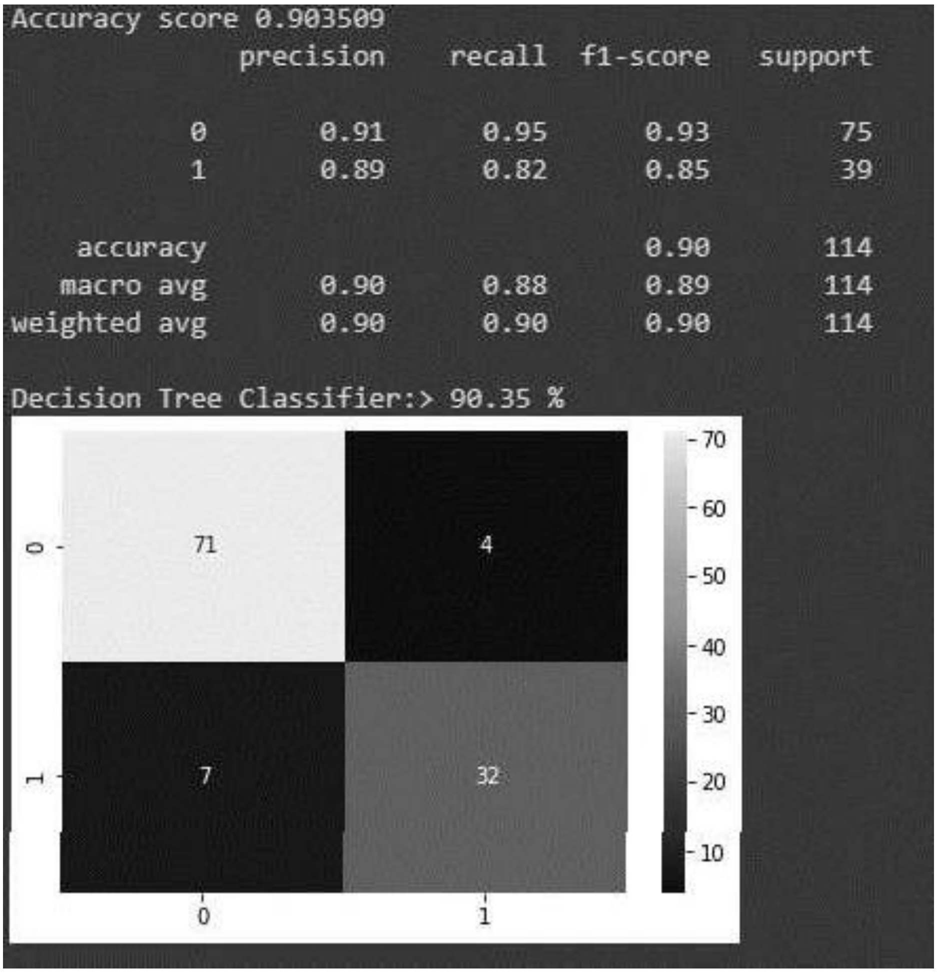
```
Accuracy score 0.991228
              precision    recall  f1-score   support

           0       1.00      0.99      0.99        75
           1       0.97      1.00      0.99        39

    accuracy                           0.99       114
   macro avg       0.99      0.99      0.99       114
weighted avg       0.99      0.99      0.99       114

SVM:> 99.12 %
```

Figure 5.2.1: SVM Result



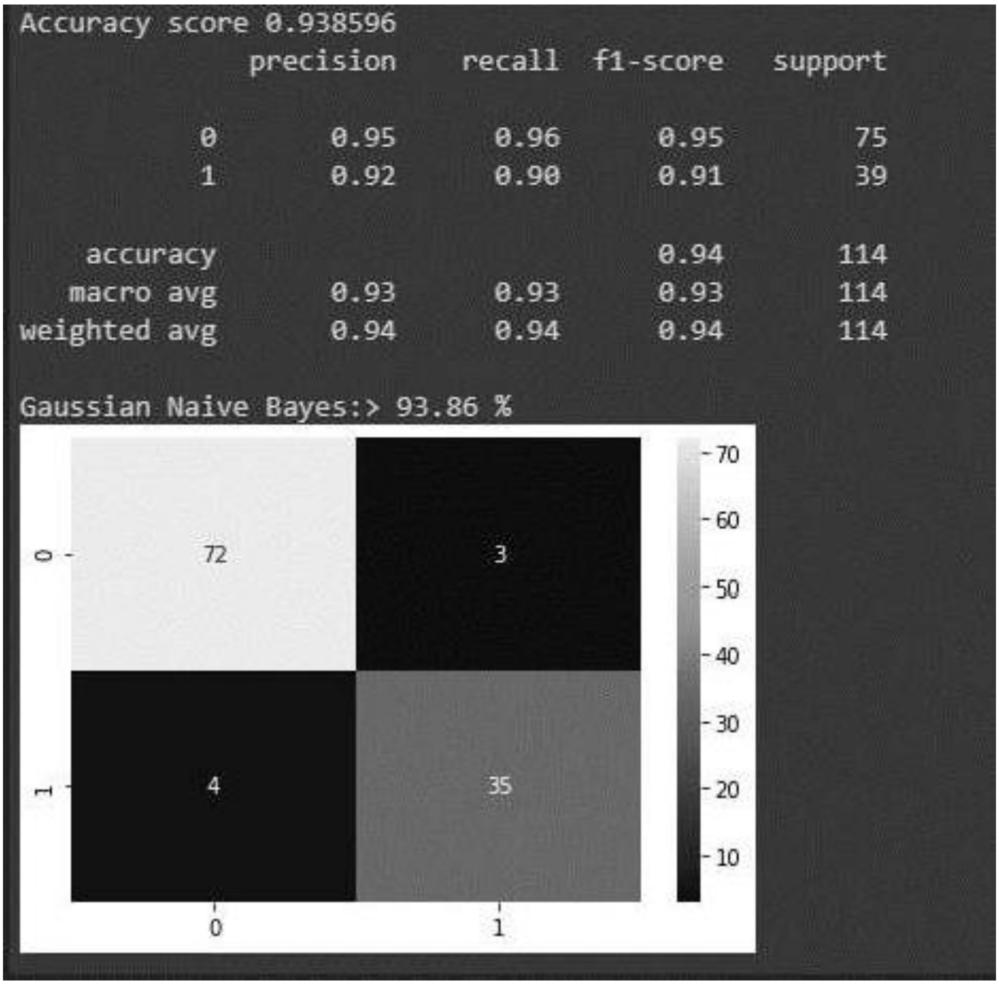
```
Accuracy score 0.903509
              precision    recall  f1-score   support

           0       0.91      0.95      0.93        75
           1       0.89      0.82      0.85        39

    accuracy                           0.90       114
   macro avg       0.90      0.88      0.89       114
weighted avg       0.90      0.90      0.90       114

Decision Tree Classifier:> 90.35 %
```

Figure 5.2.2: Decision tree Result



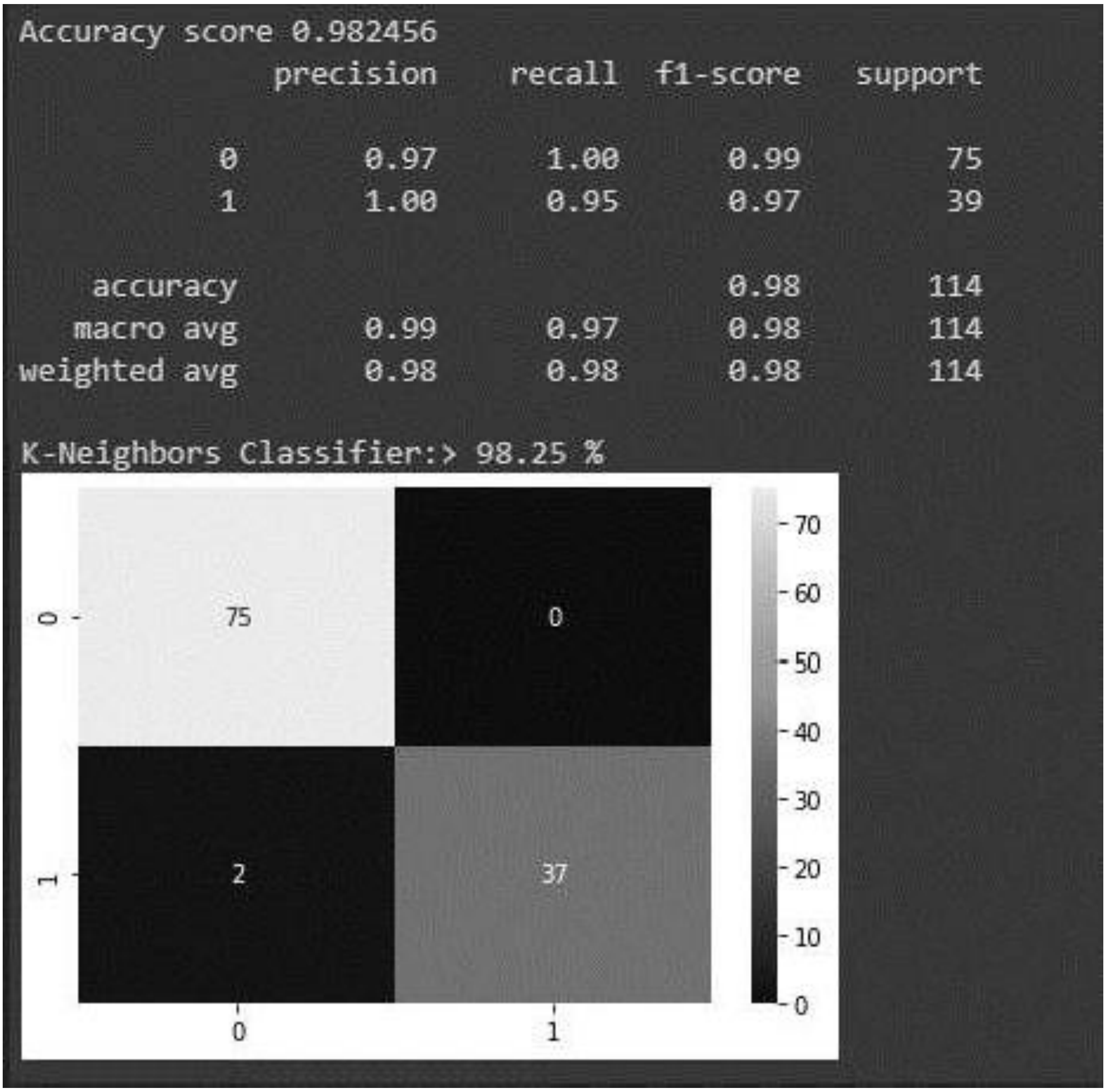
```
Accuracy score 0.938596
              precision    recall  f1-score   support

           0       0.95      0.96      0.95        75
           1       0.92      0.90      0.91        39

    accuracy                           0.94       114
   macro avg       0.93      0.93      0.93       114
weighted avg       0.94      0.94      0.94       114

Gaussian Naive Bayes:> 93.86 %
```

Figure 5.2.3: Gaussian Naïve Bayes Result



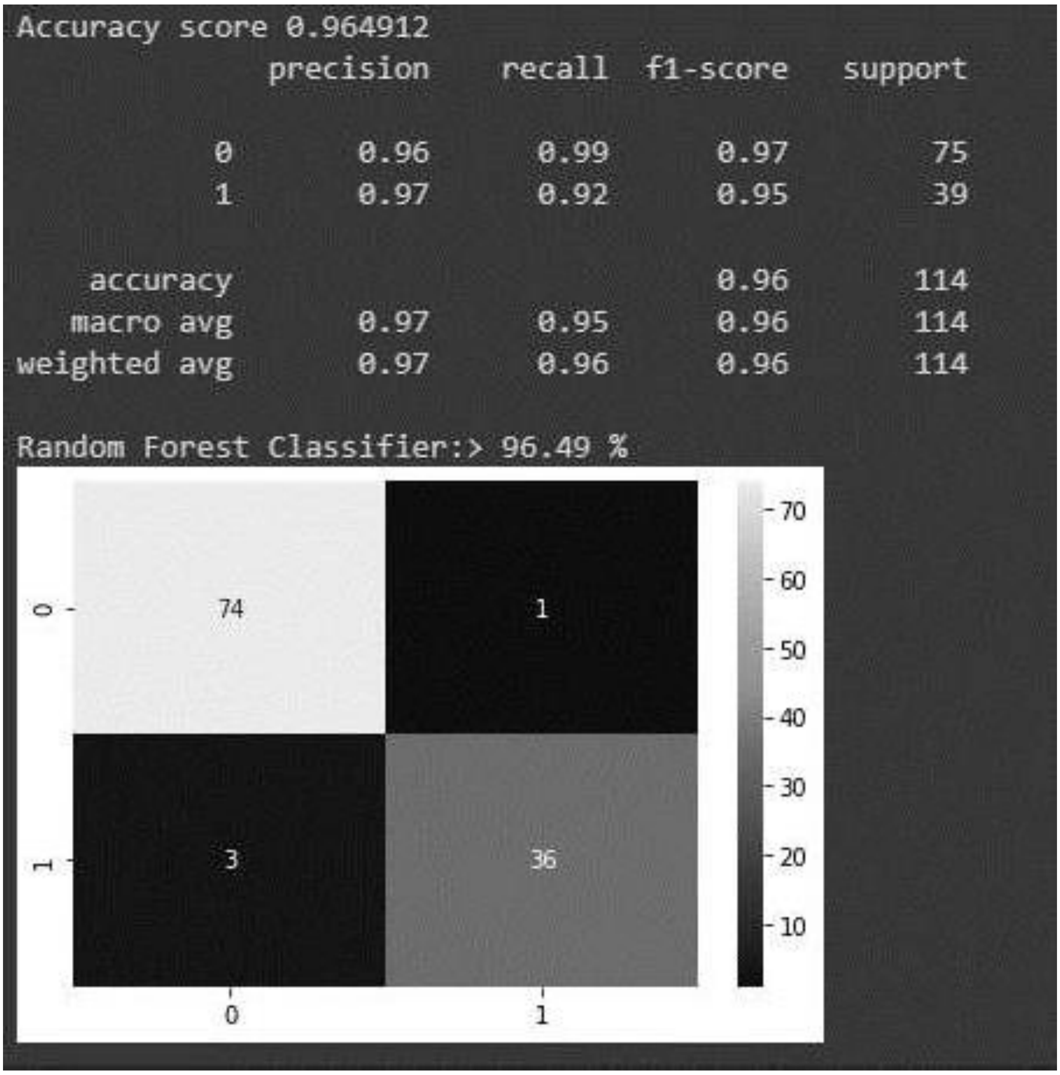
```
Accuracy score 0.982456
              precision    recall  f1-score   support

           0       0.97      1.00      0.99        75
           1       1.00      0.95      0.97        39

    accuracy                           0.98       114
   macro avg       0.99      0.97      0.98       114
weighted avg       0.98      0.98      0.98       114

K-Neighbors Classifier:> 98.25 %
```

Figure 5.2.4: KNN Result



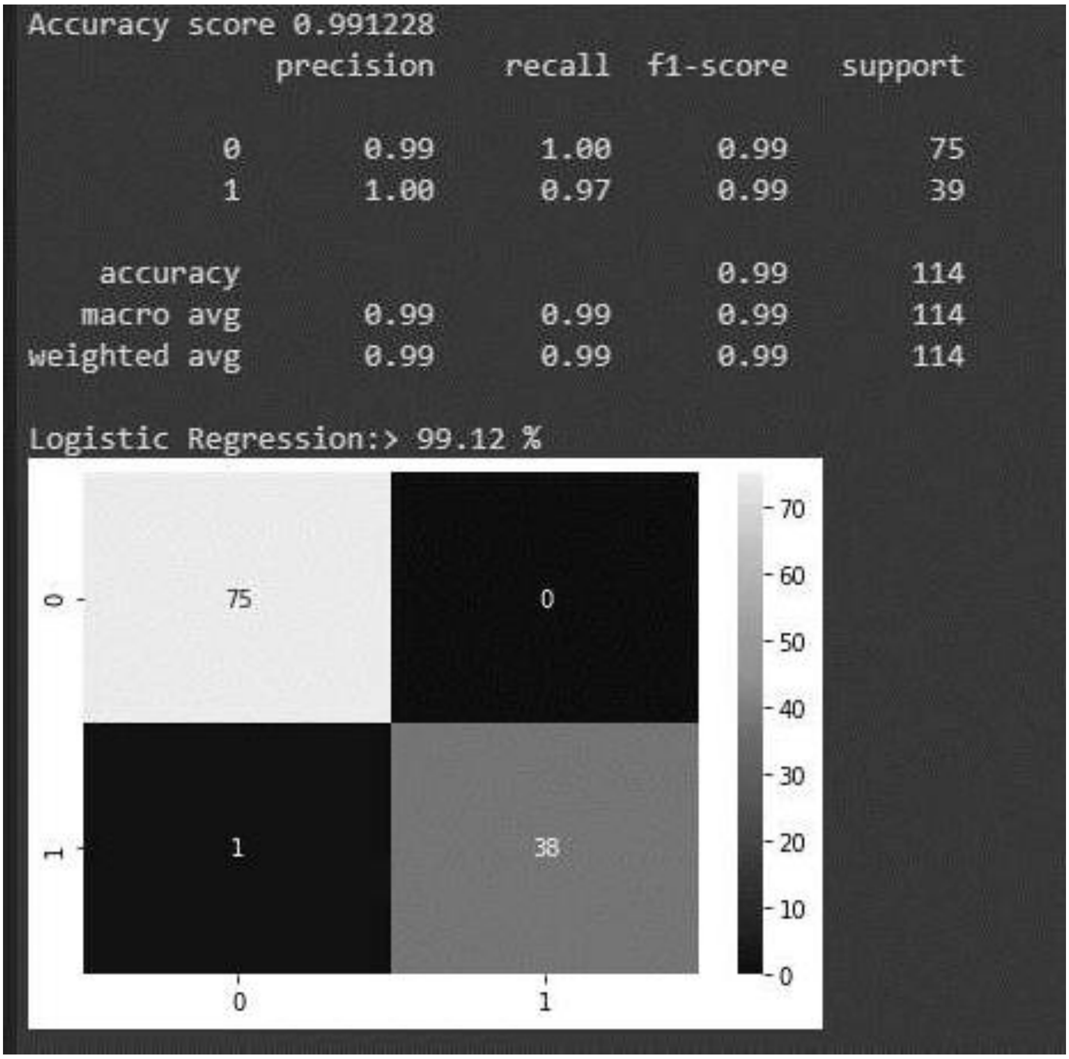
```
Accuracy score 0.964912
              precision    recall  f1-score   support

           0       0.96      0.99      0.97        75
           1       0.97      0.92      0.95        39

    accuracy                           0.96       114
   macro avg       0.97      0.95      0.96       114
weighted avg       0.97      0.96      0.96       114

Random Forest Classifier:> 96.49 %
```

Figure 5.2.5: Random forest Result



```
Accuracy score 0.991228
              precision    recall  f1-score   support

           0       0.99      1.00      0.99        75
           1       1.00      0.97      0.99        39

    accuracy                           0.99       114
   macro avg       0.99      0.99      0.99       114
weighted avg       0.99      0.99      0.99       114

Logistic Regression:> 99.12 %
```

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

Six distinct algorithms' precision is shown for inspection. The accuracy of every algorithm is over 90%. However, the accuracy of SVM and Logistic Regression methods is 99.12 percent. We conclude that SVM and Logistic Regression have the highest accuracy in predicting or detecting breast cancer when compared to the other six algorithms.

There are some other numbers that may be included to provide a better context for our study.

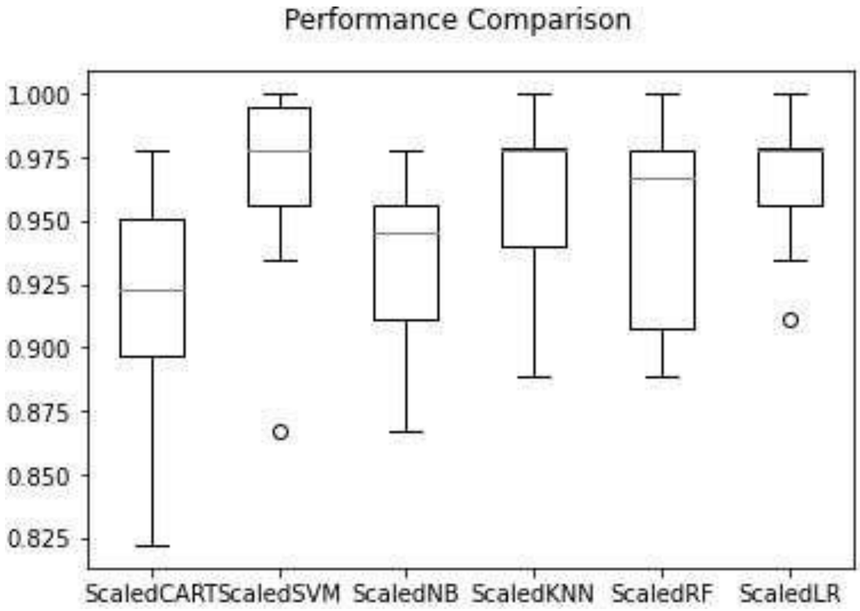
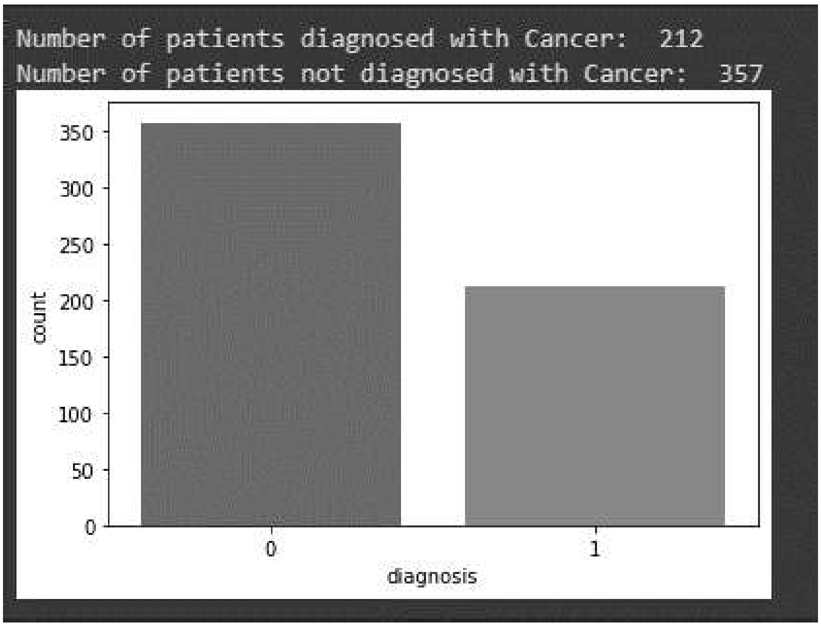


Figure 5.2.7: Performance comparison



```
Number of patients diagnosed with Cancer:  212
Number of patients not diagnosed with Cancer:  357
```

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

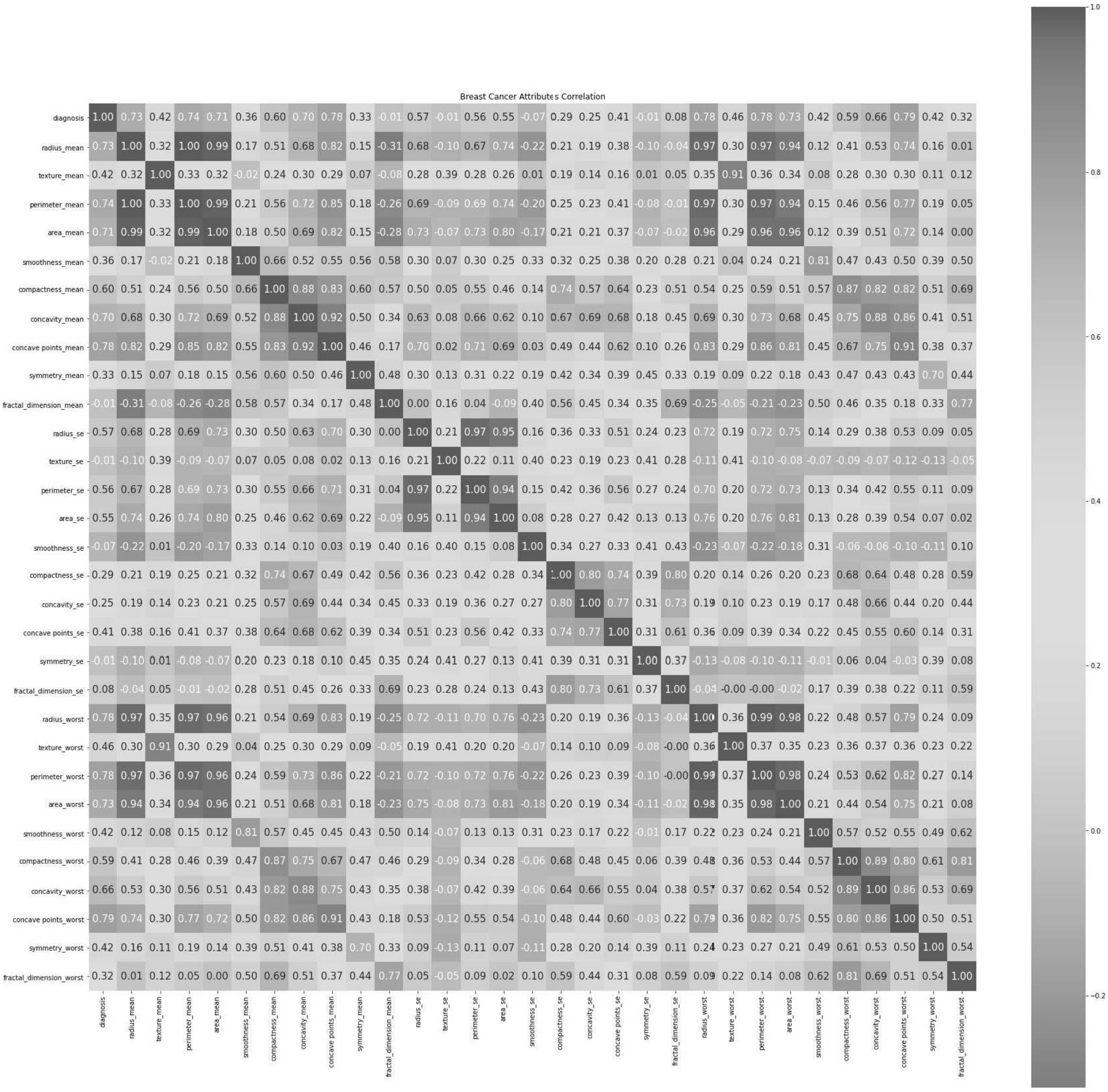


Figure 5.2.9: Breast Cancer Attributes Correlation Heatmap

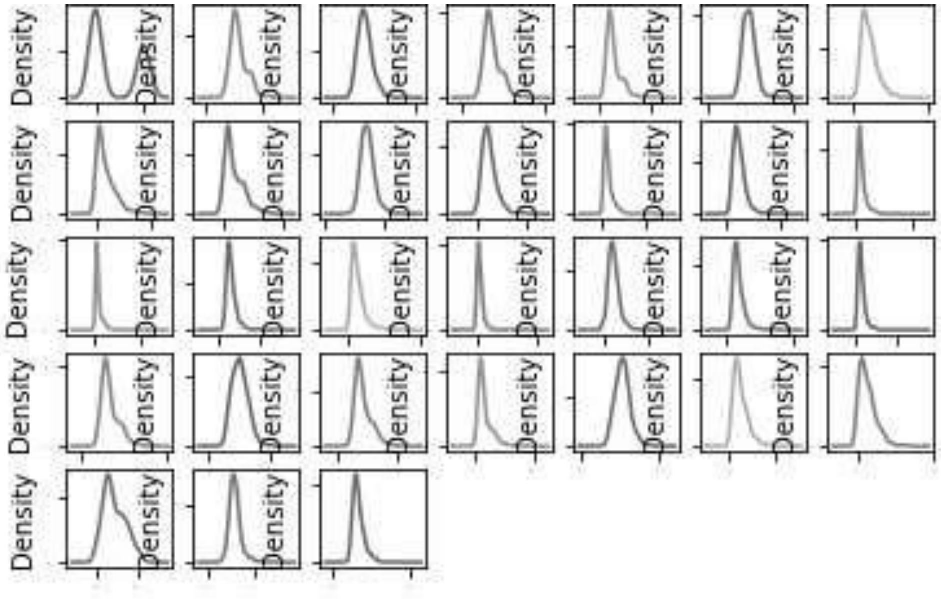


Figure 5.2.10: General Gaussian Distribution

5.3 Conclusion

In conclusion, we can state that we have made an effort to evaluate the performance of several algorithms for breast cancer detection. Support vector machine (SVM) and logistic regression (LR) techniques provide the best accuracy. Since we have only had access to 569 records, we want to expand our data and create a larger dataset with more precise numbers in the near future. We want to apply other machine learning techniques, such as support vector machine and Logistic regression algorithms, to compare the results.

CHAPTER 6

Critical Appraisal

6.1 Introduction

This section will talk about the strength, sluggishness, and breadth of my study. That will encourage the rest of the scientific community to assess this study's utility.

6.2 SWOT Analysis

Strengths, weaknesses, opportunities, and threats (SWOT) analysis is a method of strategic planning and strategic management that may be used to assess the quality of a study proposal. Scenario planning is another name for this method of assessing a given circumstance.

6.2.1 Strength

One major argument advanced by this study is that it is necessary to do this study. In reality, this study focuses on breast cancer, a major issue in human healthcare. The suggested prediction model provides a solution to this issue by allowing for the early detection of breast cancer in humans, prior to the onset of more significant symptoms. Moreover, this study establishes a bridge between the health industry and the Data Mining discipline, namely in the area of breast cancer, which is becoming increasingly important as the world becomes increasingly reliant on computer technology. From my vantage point, here is where my research really shines.

6.2.2 Weakness

The data utilized as the final dataset after clustering is of a certain kind, and this is the study's main flaw. Ultimately, this model was developed using a categorical dataset. If the dataset's input data can be partitioned into distinct groups according to the characteristics it employs, this approach may be used. To me, it seems like a flaw in the system. To data mining specialists, perhaps the most mind-boggling development is the ease with which individuals in the modern world can classify any set of numbers or other types of data. Because of this, it clearly isn't the driving force of the study.

6.2.3 Opportunity

Opportunities for this study are enormous. This approach makes it simple for physicians or diagnostic centers to determine whether or not a patient has breast cancer based on a set of standard medical indicators. From my perspective, this is a fantastic chance for anyone studying the intersection of medicine and data mining.

6.2.4 Threat

Predicting breast cancer accurately is, in my opinion, the most dangerous hazard. Additional study in the near future, however, will allow for improvement.

6.3 Conclusion

Based on a thorough study of my research's strengths, weaknesses, opportunities, and threats, I can state that the built predictive model will have a significant impact in the medical and social business fields.

CHAPTER 7

Conclusion

7.1 Conclusion

In conclusion, we are able to declare that we have made an attempt to examine the performance of a number of different algorithms for the identification of breast cancer. Techniques such as support vector machines (SVM) and logistic regression (LR) provide the highest level of accuracy. We have only had access to 569 records, therefore we are planning to increase the amount of data that we have and provide a wider dataset that has figures that are more accurate in the near future. In order to examine the differences between the two sets of findings, we plan to use other machine learning approaches, such as support vector machine and logistic regression methods.

7.2 Further Suggested Work

It is possible to do research in the future to increase the accuracy of predictions by integrating many algorithms. In addition to that, it will place an emphasis on enhancing the accuracy of categorization and will need the development of the most effective data mining strategy possible via the use of a number of different machine learning algorithms. In order to accomplish this goal, our collection of data may be put to the test using a few different approaches. It's possible that other well-known breast cancer datasets will be used in the further study that will follow. It also came to the conclusion that it would want to apply our research approaches to the study of illnesses other than breast cancer in addition to the area of breast cancer research.

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. Ensarİ, "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 MiningTechniques," 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. Ensarİ, "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. Khourdifi 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.

APPENDICES

One of the challenges we encountered was determining how to approach the project. In addition, gathering relevant data was a significant obstacle. However, we have overcome this obstacle and accomplished our goal. Our app's primary purpose is to aid in the detection and diagnosis of breast cancer. Our tool facilitates immediate recognition and comprehension of the symptoms. As a result, there is a limit on the total number of patients. To sum up, we think it's safe to say that our application, "Applying ML Algorithms for predicting Breast Cancer" would be tremendously helpful and efficient for users.

Applying ML Algorithms for predicting Breast Cancer- Saifa Sabrina Mim (ID: 213-25-041)

ORIGINALITY REPORT

20%	10%	4%	14%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to St. Ignatius High School Student Paper	12%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	4%
3	Submitted to Davenport University Student Paper	1%
4	K M McPherson. "APC of Breast Diseases: Breast Cancer - Epidemiology, risk factors, and Genetics", BMJ, 10/15/1994 Publication	1%
5	Submitted to University of St Andrews Student Paper	1%
6	publications.aston.ac.uk Internet Source	1%
7	B. Nordenskjöld. "Endocrine Therapy of Breast Cancer: A background", Acta Oncologica, 2009 Publication	<1%