Heaven's Light is Our Guide

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Rajshahi University of Engineering & Technology, Bangladesh

A Comparative Analysis of Machine Learning Classifiers by Predicting Breast Cancer.

Author

Ansar Uddin Emon
Roll No. 1503049
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology

Supervised by

Dr. Mir Md. Jahangir Kabir
Professor
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology

ACKNOWLEDGEMENT

At first, I would like to thank the Almighty Allah for giving me the opportunity and enthusiasm along the way for the completion of my thesis work.

I would like to express my sincere appreciation, gratitude, and respect to my supervisor **Dr. Mir Md. Jahangir Kabir**, Professor, Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology, Rajshahi. Throughout the year he has not only given me technical guidelines, advice and necessary documents to complete the work he has also given me continuous encouragement, advice, helps and sympathetic co-operation whenever he deemed necessary. His continuous support was the most successful tool that helped me to achieve my result. Whenever I was stuck in any complex problems or situation he was there for me at any time of the day. Without his sincere care, this work has not been materialized in the final form that it is now at the present.

I am also grateful to all the respective teachers of Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Rajshahi for good & valuable suggestions and inspirations from time to time.

Finally, I convey my thanks to my parents, friends, and well-wishers for their constant inspirations and many helpful aids throughout this work.

Date: February 25, 2021 Ansar Uddin Emon

RUET, Rajshahi

Heaven's Light is Our Guide

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Rajshahi University of Engineering & Technology, Bangladesh

CERTIFICATE

This is to certify that this thesis report entitled ***"A Comparative Analysis of Machine Learning Classifiers by Predicting Breast Cancer"*** *submitted by* ***Ansar Uddin Emon, Roll: 1503049*** *in partial fulfillment of the requirement for the award of the degree of Bachelor of Science in Computer Science & Engineering of Rajshahi University of Engineering & Technology, Bangladesh is a record of the candidate own work carried out by him under my supervision. This thesis has not been submitted for the award of any other degree.*

Supervisor

Dr. Mir Md. Jahangir Kabir
Professor
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi-6204

External Examiner

Emrana Kabir Hashi
Assistant Professor
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi-6204

ABSTRACT

Breast Cancer is a worldwide problem after lung cancer in recent days. Every year a lot of women (men also) die for this cancer. Cancer is caused by malignant tissues and so proper diagnosis of breast tissue is needed to save peoples live. For the early detection of this cancer, precise diagnosis of affected tissue is a must which can be two types i.e. malignant (harmful) or benign (not harmful). The goal of this research is to see how different machine learning classifiers perform to detect breast cancer using Wisconsin Diagnostic Breast Cancer dataset. Three known classifiers are used in this paper and after training and testing they give their individual performance. The classifiers are Logistic Regression, Decision Tree and Random Forest (with different estimators). After observing the dataset, data preprocessing and visualization of data, their prediction scores are measured by confusion matrix. Random Forest gives the better accuracy (98.25%) among other classifiers on the specific number of estimators (trees). These models slightly shows imbalanced because of higher training accuracy than their testing accuracy. This research is strongly focused on early detection of breast cancer by proper diagnosis. It also contains discussion and comparison of mentioned three classifiers with their performance.

CONTENTS

Page No.

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

1.1 Introduction

This chapter begins with the motivation behind this thesis topic. Then the literature reviews are discussed right after. Then in the proposed methodology section, the proposed system is described briefly which will solve the problem that is being dealt with. Then, in thesis contribution, contributions of the thesis are outlined. Finally, the chapter ends with a conclusion.

1.2 Motivation

Mostly female (men also) have the high risk in breast cancer. A lot of people die from lung cancer and the breast cancer is in the second one in which people suffer most [1]. Breast cancer is caused by malignant tumors in breast tissues which means the cell division occur without any control. To decrease the number of false positives and false negative predictions machine learning method can be very useful [2]. To improve the early detection these machine learning algorithms can be used. As our country is going through the digitalization, computer automated system can easily detect the cancer tumor cell. For increasing the healthcare system by using these machine learning approach, we may save some of our valuable lives. Taking response on breast cancer dataset using these machine learning algorithm, we can decide what algorithms should be used or give the better accuracy.

1.3 Problem Statement

Cancer is one of the most concerning public health problems in the 21st century. This is the source of so many headaches all around the world right now. As stated in the International Agency for Research on Cancer (IRAC), part of the World Health Organization (WHO) we know that 8.2 million people died because of cancer in 2012 and 27 million more deaths to be expected to occur until 2030 [3].

WHO also states that 57% of the newer cancer those were diagnosed in 2012 occurred in less developed regions of the world e.g. Central America and parts of Africa and Asia and 65% of

the deaths caused by cancer also occurred in these regions. As Cancer rates continue to rise along with world population, it is now the second leading cause of death among women. Women fall victims to cancer every day a great number of women are being identified with different types of cancer all over the world. Excluding skin cancer, breast cancer is the second most common cancer type for women and the mortality of breast cancer is much higher compared to any other cancer types.

In the world, mostly female suffer a lot in breast cancer. Due to the uncontrolled division of cells in breast tissue, breast cancer is caused. For the proper diagnosis of breast cancer, many process are used. Mammography or screening of breast [4] is a method to diagnose breast cancer. To check the tissues of nipple it is used by using X-rays. Usually, in the initial stage of breast cancer, cancer cannot be detected as the size of the cancer cell is very small on that stage. It is possible to diagnose cancer at the early stage through mammography, and this test takes just a few minutes [5]. It is considered that mammography or screening as the most precise method of early detection of breast cancer. But the images of digital mammogram are most of the time very hard to check due to their missing pixels or information and presence of various types of tissues [1]. For this reason, needle is used to take biopsy of the tumor and it is known as fine needle aspiration or FNA. For the biopsy of tumors, the tissue of the affected area is taken for examination under microscope. An unwanted biopsy can be caused by a false positive detection. Statistics show that only 20-30 percentages biopsy of breast tumors give the decision about cancer [6]. An actual information of cancerous tumor can be hidden because of false negative detection which can cause death of people lives. Howsoever, some tumors are in different shape, different size and some of them are like normal tissue. So, better identification methods need to be prepared to predict breast cancer [7].

1.4 Literature Review

Various new systems have been developed for the prediction of breast cancer by the improvement of medical research. There are huge research regarding this. Some significant related research are mentioned as follows.

Fatih ak [8] used Wisconsin Diagnostic Breast Cancer for the detection of breast cancer. He used data visualization and machine learning techniques including Logistic Regression, Decision Tree, k-nearest neighbors, Support Vector Machine, Naïve Bayes, Random Forest,

and Rotation Forest. He actually focused on the comparison of the machine learning algorithms through visualization and accuracy. He sliced the dataset into three types of data including all features, highly correlated data and low correlated features. In all futures included data, Logistic Regression gave the better accuracy (98.01%). He gave a clear concept by visualizing the dataset using different machine learning tools.

Kapil and Rana [9] used two dataset i.e. WDBC dataset and another breast cancer dataset which is taken from the UCI repository and used a decision tree which is modified using weight. They used Chisquare test and they kept mostly relatable features by ranking each features of dataset. For the WDBC dataset, they got about 99% accuracy on their proposed method, but for another dataset of breast cancer, they found about 85–90% accuracy.

Yue et al. [10] mainly covered on SVM, K-NNs, ANNs, and Decision Tree classifiers for the purpose of predicting breast cancer using Wisconsin Breast Cancer Diagnosis (WBCD) dataset. Authors found that, using deep belief networks (DBNs) with ANN architecture (DBNs-ANNs) is the best approach for classification. This methodology showed 99.68% accuracy, whereas for the SVM approach, the SVM method besides the two-step clustering algorithm has achieved 99.10% accuracy in classification. Using the voting technique, authors also implemented the ensemble method of SVM, J48 and Naïve Bayes classifiers. The ensemble technique gives 97.13% accuracy.

Aruna et al. [11] used Naïve Bayes, Support Vector Machine, and Decision Trees to classify a Wisconsin Breast Cancer dataset and got the best result by using Support Vector Machine (SVM) with an accuracy score of 96.99%.

Chaurasia et al. [12] compared the performance of supervised learning classifiers by using a Wisconsin Breast Cancer dataset and Naïve Bayes, SVM, Neural Networks, Decision Tree methods applied. SVM acquired the highest accuracy i.e. 96.84%.

Asri et al. [13] showed the comparative analysis of different classifiers: SVM, k-nearest neighbors, Naïve Bayes, and Decision tree (C4.5) using this dataset. The goal of this research was to classify data on the basis of efficiency and effectiveness by comparing the accuracy, precision, sensitivity, and specificity of each algorithm. On the used method SVM gives the best accuracy which is 97.13%.

Wang et al. [14] analyzed to find out the proper method of predicting breast cancer on several records of data mining. They applied support vector machine (SVM), artificial neural network (ANN), Naïve Bayes classifier, and AdaBoost Tree. Reducing the feature space was discussed, then Principle Component Analysis (PCA) was applied with the aim of reduction. Two datasets were used in evaluation section of the performance of the models they are: the Wisconsin Breast Cancer Database (1991) and Wisconsin Diagnostic Breast Cancer (1995) [15, 16]. Detailed evaluation of the models and test errors were provided by them.

Two separate datasets were used by Hasan et al. [5]. Wisconsin Diagnostic Breast Cancer dataset and SEER 2017 Breast Cancer dataset. To derive valuable characteristics, they then used Principal Component Analysis. After that, utilizing multi-layer perceptron (MLP) and convolution neural network (CNN), they categorized the reduced datasets. For both the decreased datasets, they then produced a comparative analysis of their model. The MLP model achieved an accuracy of 99.1 percent on the decreased WDBC dataset and 89.3 percent on the SEER 2017 Breast Cancer dataset, while the CNN Model achieved 96.4 percent on the decreased WDBC dataset and 88.3 percent on the SEER 2017 Breast Cancer Dataset.

1.5 Thesis Contribution & Challenges

The main contributions of the thesis are as follows:

- Finding the breast cancer dataset from UCI Machine Learning Repository. It is Wisconsin Diagnostic Breast Cancer Dataset (WDBC).
- In base paper they have applied Decision Tree, Logistic regression, SVM, Naïve Bayes Classifier and Random Forest. We have used Decision Tree, Logistic Regression and Random Forest to see these accuracy differs or not.
- At first we have visualized the patient dataset and then observed the data type, missing values and correlation of the dataset.
- In dataset some categorical values are converted into numerical value.
- Then the dataset are split into dependent and independent dataset. They are trained and tested. After training the independent variables are scaled. This is the part of data preprocessing.
- From the used algorithms, Random Forest gives the best accuracy. Increasing estimations of RF, accuracy changes are observed.

Challenges:

In this research, the main challenges is to extract the useful information from a huge pile of dataset. Preprocessing the data, visualizing the dataset with correlation, then proper use of classifiers is also a big challenge. Proper balance of dataset to train the model for overcome overfit or underfit is another difficult task.

1.6 Thesis Organization

The rest of the thesis is organized as follows:

Chapter 2 – Breast Cancer

This chapter describes the signs and symptoms, diagnosis and tests of Breast Cancer.

Chapter 3 – Research Methodologies

This chapter describes the whole methodologies including used machine learning algorithms, advantages and disadvantages/limitations of all the used algorithms. It also describes the whole dataset and visualization of some data.

Chapter 4 – Implementation Methodologies

This chapter discusses the dataset we worked with and the findings, results, and analysis of the thesis work. It also compares with other works relating to this area.

Chapter 5 – Result and Analysis

This chapter concludes the thesis, describes its limitation and shows a direction of future work.

Chapter 6 – Conclusion and Future Works

This chapter concludes the thesis, describes its limitation and shows a direction of future work.

1.7 Conclusion

Breast cancer is very common issue now-a-days in woman. By the methods and algorithms it can be detect. The process are discussed briefly in later chapter.

CHAPTER 2

Background Study

2.1 Breast Cancer

Breast cancer is a most dangerous disease in the world. Many woman die for not having proper knowledge about it. Breast cancer is caused in breast tissues.

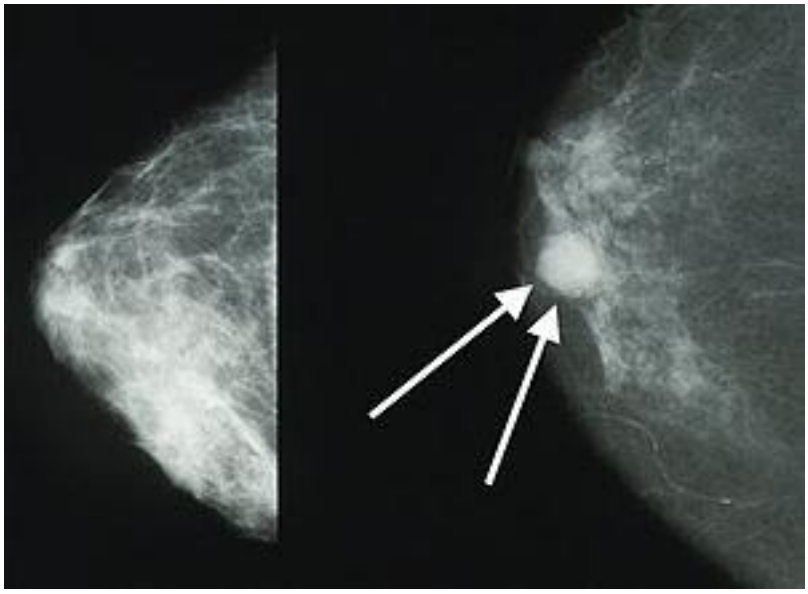

Figure 2.1: Mammogram pictures*

2.2 Breast Cancer Symptoms

Different women have different breast cancer signs. There are no signs or symptoms in certain people at all. Any early signs of cancer of the breast are-

1. New lump in the underarm or breast. Various types of lumps are breast cysts, which are soft, fluid-filled cysts, referring to sacs filled with milk that can occur during breast-feeding. But usually cysts are benign (non-cancerous).
2. Getting thicker or widening of the breast part.
3. Annoyance or creases of the skin of the breast.
4. Rosacea in the vicinity of the nipple or the breast or layered tissue.
5. Nipple squeezing or pain in the vicinity of the nipple.
6. Discharge of nipples other than breast milk, blood included.

*https://sco.wikipedia.org/wiki/File:Mammo_breast_cancer_wArrows.jpg

7. Some shift in the size of the breast or its shape.
8. Pain in every breast section.

2.3 Risk Factors

The risk factors are shown in the below:

1. Being woman
2. Obesity
3. Ignorance
4. Lack of exercise
5. Alcohol
6. During menopause, hormone replacement therapy
7. Ionization radiation
8. First menstruation at the early stage
9. Having children late in life or not at all
10. Older age
11. Prior breast cancer
12. Family history of breast cancer etc.

2.4 Diagnostic Method

Mainly breast cancer is diagnosed in two ways.

1. **Biopsy:**

   In this method a tissue of the target area is taken. Then it is sent to a pathologist. The pathologist or doctor observe it under the microscope. By needle fluid of the target area are taken. It is also called fine needle aspiration biopsy (FNA).

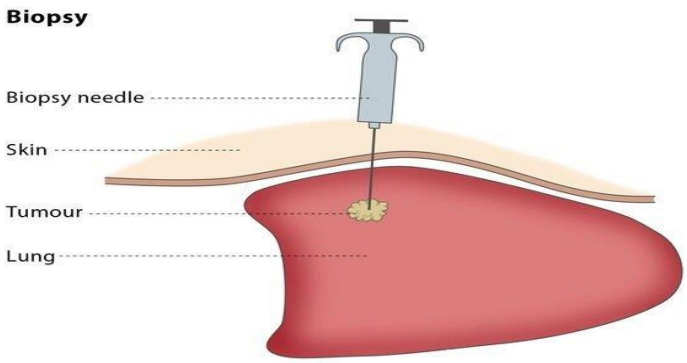

Figure 2.2: Biopsy using needle[17]

2. **Mammography:**

It is actually a process of X-ray with low voltage. To get the better image screening mammography is used. It plays a very important role for the detection of breast cancer early.

Figure 2.3: A mammography machine[18]

2.4 Treatment

There are some world-wide used treatment are described in the following:

1. **Surgery:**

The target area or affected area with cancerous cell are permanently cut by surgeon.

2. **Radiation Therapy:**

It is also known as RT, RTx or XRT. By ionization, it kills the malignant tumor in the affected area.

3. **Chemotherapy:**

It is known as CTX or CTx. It is the process of using anti cancerous drugs to kill the cancerous cell. But it has many side-effect too. It also damages the good cell, hair follicles tissue, bone marrow etc.

4. **Hormonal Therapy:**

It is used to such type of cancer which is caused by hormones. Selective Estrogen Response Modulator is largely used for the breast cancer treatment.

2.5 Conclusion

This chapter describes the breast cancer, symptoms and treatment of the disease. Mostly woman should have the knowledge about this as it is very deadly in recent time.

CHAPTER 3

Research Methodologies

3.1 Introduction

In this chapter we would describe our methodologies and workflow by which we can observe the dataset. It also covers the machine learning algorithms which are used in this research.

3.2 Machine Learning

Machine learning is the subset of Artificial intelligence. It means some algorithms or approaches by which any prediction is gained. For decision making, learning from experience is the main part of machine learning by which it can give proper detection. Some examples of machine learning are medical diagnosis, stock market prediction, house price prediction, image processing, regression etc.

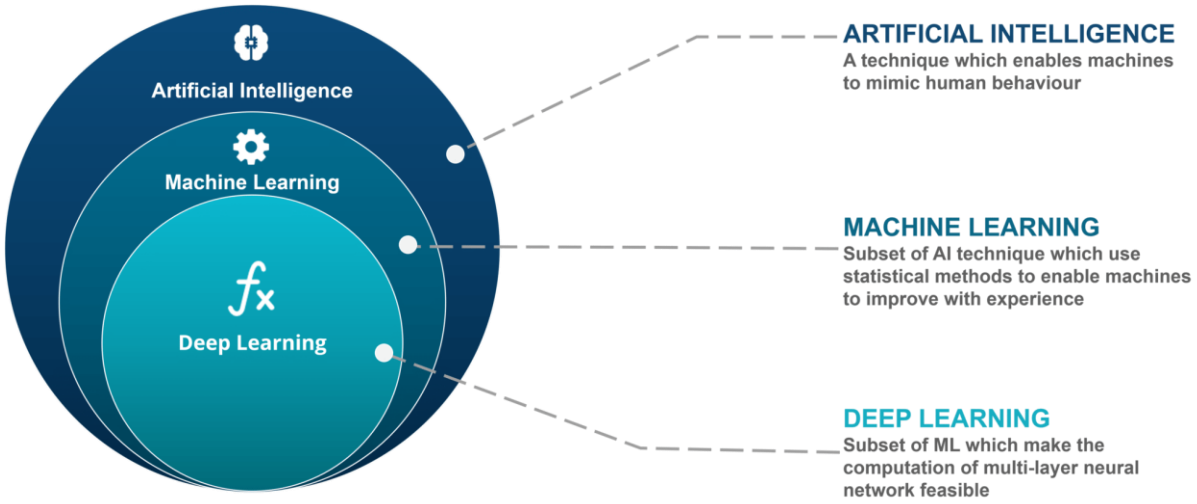

Figure 3.1: A snapshot of machine learning[19]

3.3 Machine Learning Methods

Some methods of machine learning are described in the following.

3.3.1 Supervised Machine Learning

It is the process of learning dataset with label. Suppose in a dataset of breast cancer tumor, the class label is “Benign” or “Malignant”. If we trained the dataset with existing

variables, then it automatically generates a function with dependent and independent variables. If new value comes, then from the experience of learning it will give prediction. For this reason, it is called supervised machine learning.

3.3.2 Unsupervised Machine Learning

Unsupervised learning method is the method where the dataset has no class attribute or label. If the dataset is trained, then it works on a default function but it doesn't find out the right values. It can describe the dataset and interpret the relationship in whole dataset.

3.3.3 Semi-supervised Machine Learning

It is the mixture of labeled data and unlabeled data. If so many labeled data are trained the unlabeled data cause effects on the labeled data. It is actually used to see the skill or performance of trained label data. It is the medium level between supervised and unsupervised learning.

3.3.4 Reinforcement Machine Learning

It is known as learning from experience. By trial and error method it works and the prediction is found delay for this. It interacts with surroundings and produces error or reward from this.

3.4 Mostly Used Machine Learning Algorithms

Various types of machine learning algorithms are used to predict anything. Some mostly use algorithms are mentioned in the following:

(i) Logistic Regression
(ii) Linear Regression
(iii) Decision Tree
(iv) SVM
(v) Naive Bayes
(vi) kNN
(vii) K-Means

(viii) Random Forest

(ix) Dimensionality Reduction Algorithms

(x) Gradient Boosting algorithms

1. GBM
2. XGBoost
3. LightGBM
4. CatBoost

3.5 Random Forest Classifier

Forest is made by lots of trees. Random Forest is also made by a lot of Decision Trees. Most of the machine learning problems are based on regression and classification. Random forest can be used both classification and regression problems. This chapter contains the details of random forest classifier algorithm. It is the most flexible and mostly used algorithm in machine learning platform.

3.5.1 Introduction

Because of its simplicity and diversity, it is now one of the most used algorithms (it can be used for both classification and regression tasks)[20]. It gives better result even without hyper-parameter tuning.

3.5.2 Hyper-parameter Tuning

Hyperparameters are referred to as parameters that describe the model architecture, so this method of looking for the optimal model architecture is referred to as hyperparameter tuning. The methods of hyperparameter tuning contribute towards how we sample potential candidates for model architecture from the domain of feasible hyperparameter values. This is sometimes pointed to as "finding" for the optimal values in the hyperparameter domain. When building Random Forest model, two important hyperparameters to consider,they are:

(i) Finding the optimum amount of estimators (i.e. decision trees).

(ii) Finding the maximum allowable depth for each decision tree.

In the following visualization, the xx and yy dimensions represent two hyperparameters, and the zz dimension represents the model's score [21].

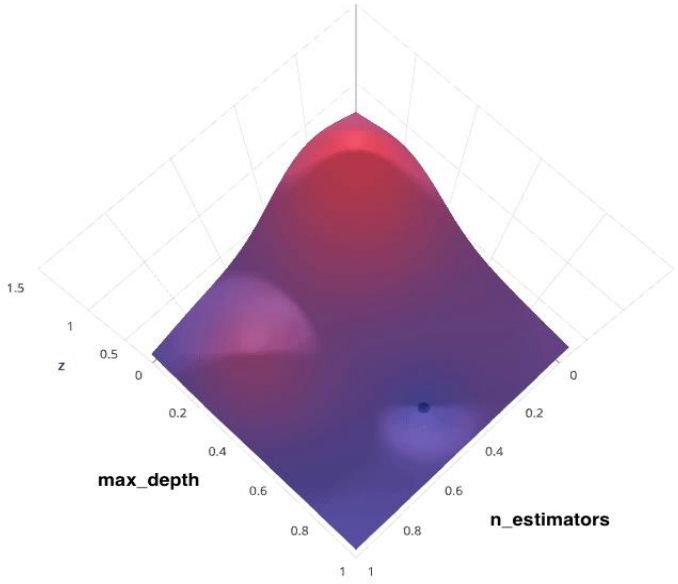

Figure 3.2: A Visualization of Hyper-parameter Tuning of Random Forest [21]

3.5.3 Construction of Random Forest

It is a supervised learning model which is a combination of different decision trees. Every decision tree gives a result and by merging the prediction score random forest gives a better result. A snapshot of how random forest works is given in the below:

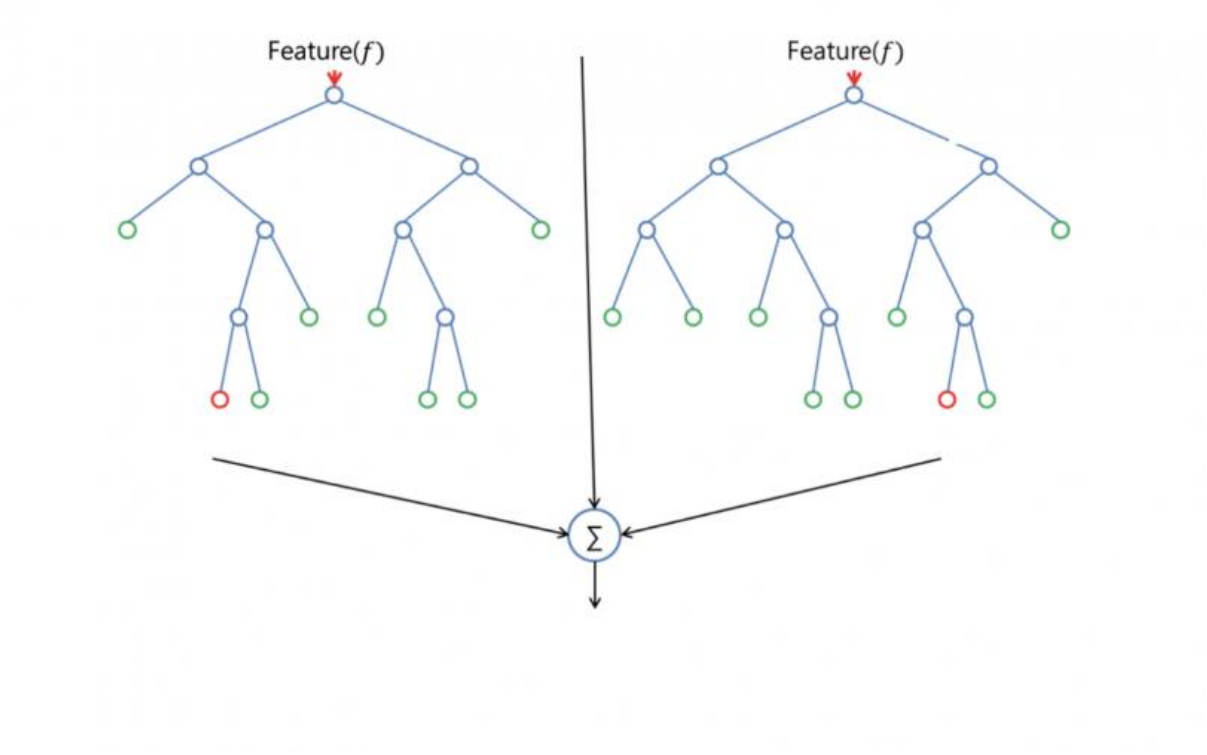

Figure 3.3: Construction of Random Forest using decision trees[20]

In the above picture we can see that there are two trees. After taking prediction score from different decision trees it gives more stable prediction score by adding them together.

3.5.4 Working Flow

The workflow of random forest is given below:

i. Selection of K data points from the training package randomly.
ii. Generating the decision trees from these K data points.
iii. Choosing the number of N-trees from the produced trees, and repeat steps (i) and (ii).
iv. From the N-tree, which forecasts the group to which the data points refer for a new data point, and specifies the new data point with the greatest probability across the group [22].

3.5.5 Difference between Decision Tree and Random Forest

There are a huge difference between decision tree and random forest and there lies many important logics behind it.

(i) A decision tree is a simple tree which gives a decision and possible consequences. A random forest is a technique of ensemble learning that works by creating a mixture of decision trees.

(ii) There is a possibility of overfitting but Random Forest reduces the risk of overfitting.

(iii) Decision Trees gives less accurate result. Random Forest gives it more accurately.

(iv) Decision Trees is simpler to understand, interpret than Random Forest.

3.5.6 Advantages of Random Forest

One of the most important advantages of Random Forest is that it is used both classification and regression problems. Because of it's using of default hyperparameters it gives better accuracy most of the time. If there are more trees in the forest, overfitting will not be occurred. It also handle multi-collinear data in dataset.

3.5.7 Disadvantages of Random Forest

Though it reduces overfitting problems by the presence of lot of trees, it also takes time for the large number of trees. Random Forest is fast to train but slow to prediction because of more decision trees.

3.5.8 Conclusion

The algorithm is also a great choice for anyone who needs to develop a model quickly. On top of that, it provides a pretty good indicator of the importance in assigned features. Random forests are also very hard to beat performance wise. Surely, we can still find a model that can do better, such as, for example, a neural network, although these typically take longer to create, although they can manage a lot of different types of features, such as binary, categorical and numerical. Eventually, random forest is an easy, simple and versatile tool (mostly), but not without certain limitations. We have used random forest classifier based on entropy or information gain for our research purposes.

3.6 Decision Tree

3.6.1 Introduction

Decision tree (DT) provides powerful techniques for classification and prediction [23], [24]. There are several algorithms to build DT model. DTs are robust classification algorithms, which are becoming popular with the growth of data mining in the information systems field. DT-based classification algorithms have tree structures consisting of nodes (or leaves), and branches. The tree structure is constructed based on a set of decision rules applied in a certain order [25]. It is a Supervised Machine Learning where the data is continuously split according to a certain parameter [26].

3.6.2 Construction

A Decision Tree constructed by nodes, edges, leaf nodes. They are also called decision nodes, chance nodes and end nodes. The purpose of DT building is to search for a set of decision rules to predict an outcome from a set of input variables. There are two main phases of the DT induction process: the growth phase and the pruning phase. The growth phase involves a recursive partitioning of the training data resulting in a DT where decision trees have a natural "if", "then", "else" construction that makes it fit easily into a programmatic structure [8]. We can get a specific idea from the picture below:

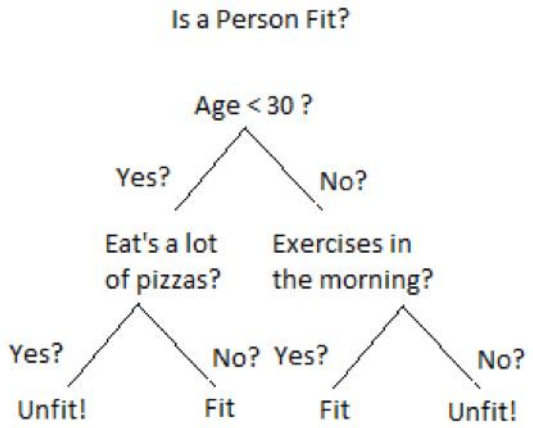

Figure 3.4: An Idea of Decision Tree[26]

In the picture 3.4 it will give decision about a person in fit or unfit. In every node, there is a parameter by which a decision is taken. Every node is come from its ancestor node. If there

will be no node in the last node then it is considered as leaf node. A leaf node actually contains the final prediction result of a tree.

3.6.3 Categories

Various decision tree algorithms are available to guide the classification of data, including ID3, C4.5, C5, CART and CHAID [27]. There are two main types of DTs.

(i) **Classification Trees:** This type of trees are built by binary recursive partitioning. In figure 3.4 we can see that a person is fit or unfit which is categorical or discrete values. This type of trees do not give outcomes in continuous decision. They give the decision which are yes/no type or will occur or not occur.

(ii) **Regression Trees:** In this trees the target value is in continuous form. That means any independent value there is a predicted value which is not in discrete format but it will give a continuous value. Suppose, the price of the house prediction is in continuous value.

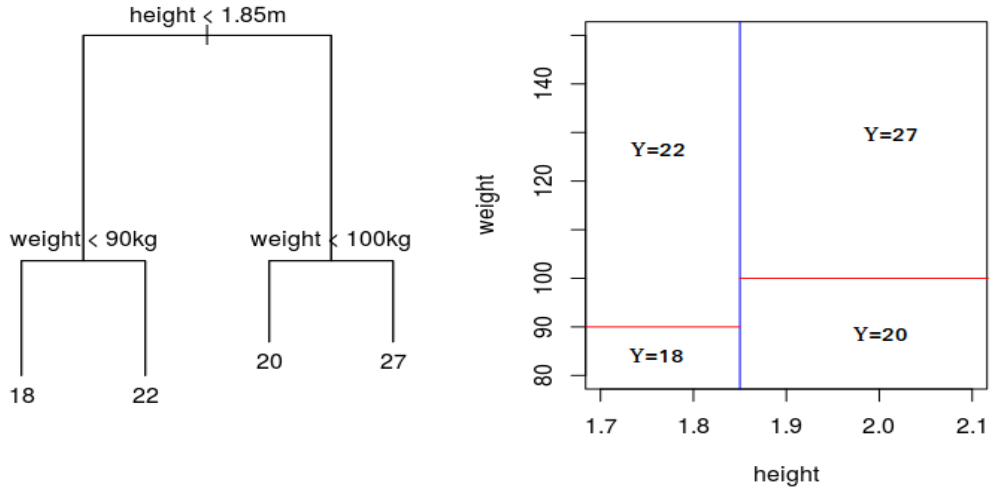

Figure 3.5: An Example of Regression Tree [26]

3.6.4 Construction of Decision Tree (DT)

A decision tree is constructed by recursively partitioning the feature space of the training set. The objective is to find a set of decision rules that naturally partition the feature space to provide an informative and robust hierarchical classification model [28]. The training dataset is divided into many subset and gradually a decision tree is created. At last a full

decision tree is returned in default. By divide and conquer method a DT is created. It follows mainly in the way of:

(i) First of all, selecting a root node with large information gain, it creates probable child nodes.

(ii) Splitting the nodes with same class of root node, leaf nodes are created which must be pure. But sometimes leaf node may have impurity too.

(iii) Repeating the process of nodes splitting, if the nodes of same classes are attended, then recursion of splitting tree process will stop. Then a branch of a tree will be created.

3.6.5 Gini Index and Randomness

At the very beginning of this algorithm, it is essential to choose the best attribute and place it at the top on tree figure and then split the tree. Gini index and information gain are two methods for the selection of features. Randomness or uncertainty of feature x is defined as entropy and can be calculated as follows [8]:

$$H(x) = Ex[I(x)] = -\sum P(x)logp(x)$$

Entropy values for each variable are calculated, and by subtracting these values from one, information values can be obtained. A larger data gain boosts an attribute and positions it on top of the tree[8].

Gini index is a measure of how often a randomly chosen element would be incorrectly identified. A lower Gini index score, thus, means better attributes. We can find the Gini index with the following equations [8]:

$$G = \sum pi * (1 - pi) for\ i = 1, \ldots . n$$

We have used randomness or entropy for our classification of dataset.

3.6.6 Advantages of DT

A DT is very inexpensive to construct. It is very simple to construct and removes unimportant features. It gives better accuracy most of the times. Decision trees are capable of handling datasets that may have missing values [29]. Decision trees are capable of handling

datasets that may have errors. It is used for its less complexity and easy to decode a single tree. Adding alternative leaf node classifiers to decision tree models can be easy also[28].

3.6.7 Disadvantages of DT

Its main problem is that overfitting. Having large number of levels of features, DT seems biased to this. Even small changes in any features can't change a DT very much. A large DT performs slowly. For this a group of decision tree that is Random Forest is used.

3.6.8 Conclusion

Decision Trees are known to be one of the most common methods for classifier representation. The topic of growing a decision tree from available data has been discussed by researchers from various fields such as analytics, machine learning, pattern recognition, and Data Mining.

3.7 Logistic Regression

3.7.1 Introduction

Regression is the relationship between independent (explanatory) variables and dependent (response) variables. If we consider the explanatory variables as $X_1, X_2, X_3 \ldots\ldots X_n$ and the response variables as y then regression gives the strength of response variable and explanatory variables. The various regression models differ mainly through the type of response variables (continuous, binary, categorical, or counts) and the different kinds of covariates, which can also be continuous, binary, or categorical. In the logistic regression model, the dependent variable is in binary form [30].

3.7.2 An Overview

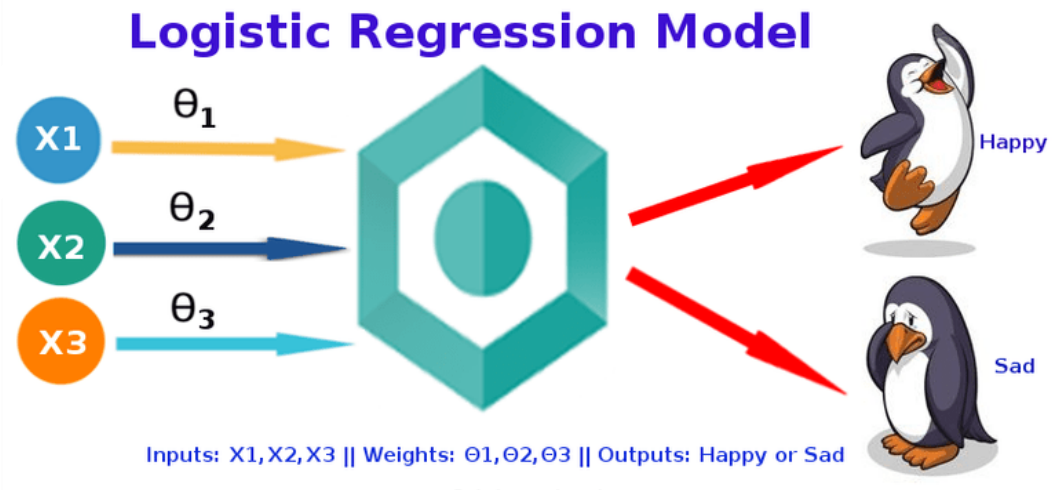

Figure 3.6: Model Visualization of Logistic Regression[31]

Suppose depending on its everyday habits, Penguin needs to know how likely it is that it will be happy. When the penguin wants to create a model of logistic regression to predict happiness based on its daily operations. Both the fun and sad activities are important for the penguin. These operations are considered as the Input parameters in machine learning terms (features) [31].

Let's consider a scenario where we need to classify whether a tumor is malignant or not. If we use linear regression for this problem, then we must have set a threshold for the prediction. Suppose, if the real class is malignant, the expected continuous score is 0.4, and the threshold

score is 0.5, the data point would be labeled as non-malignant, which will result in severe immediate effects.

3.7.3 Logistic Regression Hypothesis

The logistic regression classifier can be derived by analogy to the linear regression hypothesis which is [32]:

$$h(\theta) = \theta^T x$$

After derivation the logistic regression hypothesis will be:

$$h(\theta) = g(\theta^T x)$$

$$g(z) = \frac{1}{1 + e^{-z}}$$

Here $g(z)$ is known as sigmoid function. So logistic regression hypothesis will be:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

Here "z" takes the real values and $g(z)$ is between 0 and 1.

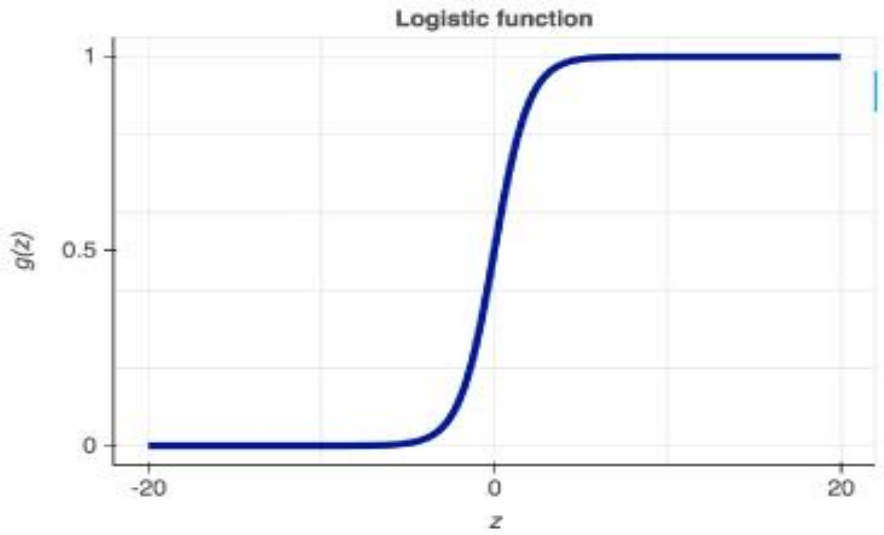

Figure 3.7: A Sigmoid Function[32]

3.7.4 Method

The logistic regression theorem serves to restrict the cost function to between 0 and 1. It is thus not represented by linear functions so it may have a value greater than 1 or less than 0, which, according to the logistic regression theorem, is not feasible.

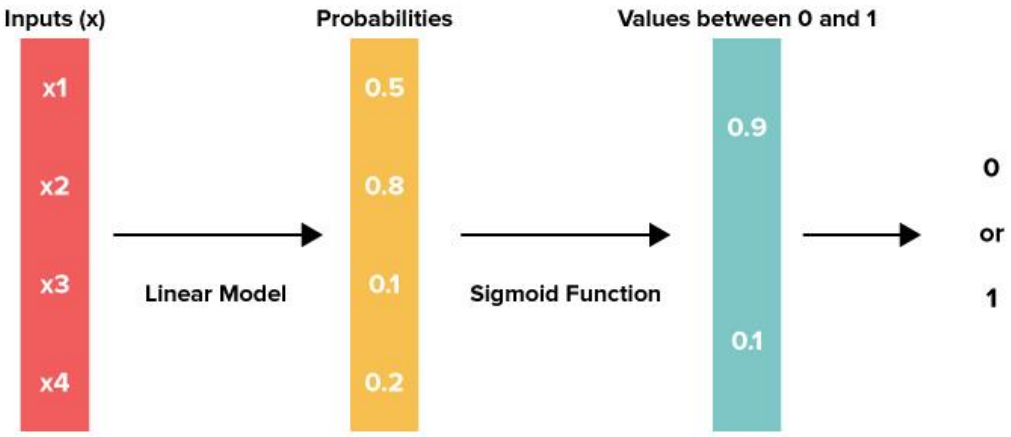

Figure 3.8: How Logistic Regression Works [32]

3.7.5 Cost Function

Cost function for Logistic Regression are [32]:

(i) $\text{Cost}(h_\theta(x),y) = -\log(h_\theta(x))$ if $y = 1$

(ii) $\text{Cost}(h_\theta(x),y) = -\log(1-h_\theta(x))$ if $y = 0$

The above functions can be written together as:

$$J(\theta) = -\frac{1}{m}\sum[y^{(i)}\log(h\theta(x(i))) + (1-y^{(i)})\log(1-h\theta(x(i)))]$$

After minimizing the cost function using gradient descent we can get,

$$\theta_j := \theta_j - \frac{\alpha}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}$$

After solving the derivative part,

$$\theta_j := \theta_j - \frac{\alpha}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}$$

3.7.6 Categories of Logistic Regression

1. Binary Logistic Regression

There are only 2 potential outcomes for the definitional answer. Example: Benign or Malignant.

2. Multinomial Logistic Regression

Without ordering three or more groups. Example: determining which meal is more preferable

(Non-Veg, Veg or Vegan).

3. Ordinal Logistic Regression

Three or more categories with ordering. Example: Movie rating from 1 to 5.

3.7.7 Difference between Logistic Regression and Linear Regression

Difference between Logistic Regression and Linear Regression are in the following: [32]

(i) The result (dependent variable) in linear regression is continuous. It may have any of an unlimited number of values that are feasible. But the response (dependent variable) in logistic regression has only a small range of possible values.

(ii) When the output variable is continuous, linear regression is used. For example, status of weather, number of hours, etc. But where the outcome variable is categorical in definition, logistic regression is used. Yes/no, benign/malignant, red/green/blue, true/false, etc, for example.

(iii) The data points can be correlated with each other in Linear Regression. The explanatory data points cannot be correlated with each other in logistic regression (no multi-collinearity).

(iv) Linear regression provides an equation of the form that is $Y = mX + c$, it contains the degree of 1 in the equation. Logistic regression gives an equation which is of the form $Y = e^{x} + e^{-x}$

(v) The coefficient representation of explanatory variables is very simple in linear regression (i.e. keeping all other variables constant, the predictor variables is assumed to increase/decrease with a unit rise in this variable). Depending on the family (binomial, Poisson, etc.) and relation (log, logit, inverse-log, etc.) we use in logistic regression, the explanation is distinct.

(vi) In order to reduce the errors and obtain the absolute best fit, linear regression uses the usual least square method, while logistic regression uses the maximum probability method to arrive at the approach. Just the reverse is logistic regression. Using the function of logistic loss allows an asymptotic constant to be penalized for significant errors.

3.7.8 Advantages of Logistic Regression

(1) Because of the logistic function on which the model is based, the logistic model is familiar, provides the following:

(i) Estimates that must lie in the range between zero and one.

(ii) An enticing S-shaped definition of the cumulative impact on disease risk of many risk factors [30].

(2) Not only does the logistic regression model function as a model of classification, but it also provides us with probability. This is a major improvement over other models where the final classification can only be produced. Compared to 51 percent, understanding that an example has a 99 percent chance for a class makes a major difference. When the dataset is linearly separable, Logistic Regression works effectively.

(3) Logistic Regression not only provides a metric of how significant an indicator is (coefficient size), but also of its correlation path (positive or negative). We see that it is simpler to apply, perceive and train logistic regression very efficiently.

3.7.9 Disadvantages of Logistic Regression

(1) Due to full isolation, logistic regression will suffer. The logistic regression model can no longer be learned if there is an attribute that can completely distinguish the 2 groups. This is because there will be no convergence of the weight for that attribute, because the optimum weight would be infinity. This is a little disappointing too, as such a feature is really very beneficial. But since we have a clear rule that distinguishes all classes, we do not require machine learning. By applying penalty system of the weights or specifying a prior probability distribution of weights, the problem of total separation can be solved.

(2) Logistic regression is less vulnerable to overfitting, but in high-dimensional datasets it can overfit, and in that situation, in such cases, normalization strategies should be known to prevent overfitting.

(3) Often, researchers misinterpret a prediction variable for the effect variable or vice versa.[33].

(4) It does not inherently mean that a causal association occurs to establish a clinically meaningful relationship between the indicator and predicted variables. [33].

(5) There may be other predictor variables that were not included in the model that have equal or greater impact on the outcome variable [33].

(6) The predictive ability of a model does not apply to data outside the range of the data from which the model was derived [33].

3.7.10 Conclusion

Logistic regression can be a powerful statistical procedure when used appropriately [34]. It has the inherent benefit of still making biologically valid predictions, and it also forecasts nearer to the results of most situations. Whenever the findings are viewed as percentages, logistic and non-linear regression should be used, however linear models which provide appropriate fit quality under some conditions.

3.8 Methodologies

In this section we will discuss about the methodologies and workflow of our total research. For implementation of any research approach, we have to go through a rough sheet i.e. called methodology. The snapshot of methodology is given in the below:

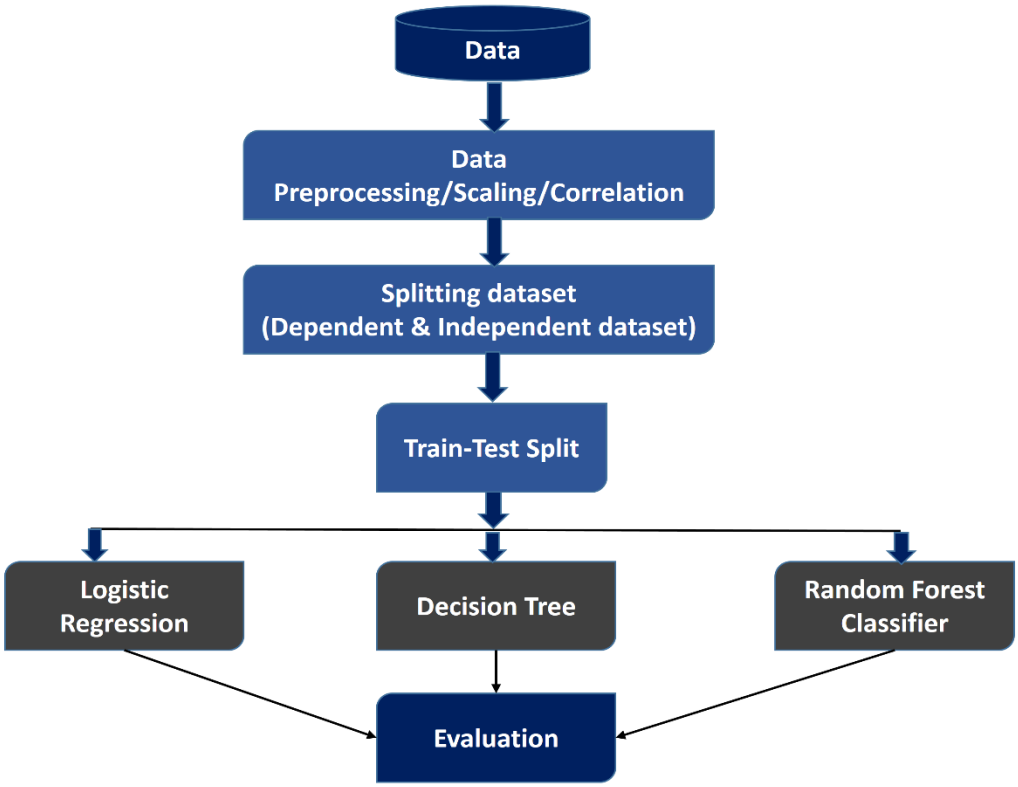

Figure 3.9: Methodology

In the 3.9 figure, first of all, data is collected from a renowned source. Then we have to do analysis and exploration on the dataset. After preparing the data, they are fed into the algorithms (Logistic Regression, Decision Tree, Random Forest Classifier). Then collecting individual performance we can make a decision from it.

3.8.1 Workflow

So the workflow of the methodology are given in the below:

(i) Collecting the data from a renowned machine learning repository.

(ii) Dataset should be observed and to see whether it is in supervised or unsupervised learning.

(iii) Then dataset has to be observed carefully. Data preprocessing (Data Cleaning, Data Transformation, Missing Value Handling, Categorical Value Handling, Scaling etc. will be done with this dataset.

(iv) From analyzing (ii) and (iii) we have to implement the dataset into different machine learning algorithms. On the methodology, three different machine learning algorithms are mentioned. They are: Logistic regression, Decision Tree and Random Forest.

(v) After going through the algorithms, performance of each algorithms are measured on training and testing accuracy. Here confusion matrix can be used to measure the accuracy.

(vi) Then we can draw a conclusion from this methodology.

3.9 Dataset

We have used two publicly available datasets in order to evaluate our breast cancer diagnosis models. The name of the dataset is Wisconsin Diagnostic Breast Cancer Dataset (WDBC) [5]. This dataset is mainly collected from the University of Wisconsin Hospital in 1995. This dataset Contains 569 samples and 32 patient attributes which includes patient ID, 30 attributes about tumor diagnosis and One diagnosis result saying if the tumor is benign or malignant [16], [35].

3.9.1 Attribute Information

The features are taken from the image of breast mass with fine needle aspirate (FNA). This dataset contains the measurement of cell nuclei showed following information:

Attributes[36]

1) ID number

2) Diagnosis (M = malignant, B = benign)

And (3 to 32) columns contains the information of FNA image.

The dataset contains 32 parameters. All parameters can be useful to classify cancer; if these parameters have relatively large values, it can be a sign of malignant tissue. The first parameter is ID, and it is a number that is used for identification [16].

Table 3.2 Attributes of Wisconsin Breast Cancer Dataset[8]

1	Id	9	Symmetry Mean	17	Smoothness Se	25	Perimeter Worst
2	diagnosis	10	concavity mean	18	compactness se	26	area worst
3	radius mean	11	concave points mean	19	concavity se	27	smoothness worst
4	texture mean	12	fractal dimension mean	20	concave points se	28	compactness worst
5	perimeter mean	13	radius se	21	symmetry se	29	concavity worst

6	area mean	14	texture se	22	fractal dimension se	30	concave points worst
7	smoothness mean	15	perimeter se	23	radius worst	31	symmetry worst
8	compactness mean	16	area se	24	texture worst	32	fractal dimension worst

The second parameter is the diagnosis of membranes, of which there are two diagnoses for tissue: malignant and benign. For different cancer types, it is necessary to determine the correct diagnosis of tissue in case both membranes have different treatments. After these two, estimated means, standard errors, and radius means indicate a range between the center and point on the perimeter. Radius se shows the estimated standard error [8].

For the approximate range, the worst radius has the center's highest value. Knowing the distance between the middle and the point is critical since operation relies on the scale. For massive tumors, there is no possibility of operation. The standard deviation of the gray-scale values is the Texture Mean [8].

Texture se represents the standard error of the calculated standard deviation for gray-scale values.

The highest mean standard deviation value for gray-scale values is displayed as the texture worst. Grayscale is widely used to find the position of the tumor, and the standard deviation is important to find the data variance and to illustrate how the numbers can be scattered. Perimeter mean represents the mean value for the core tumor, while standard error of the mean represents the core tumor described as perimeter se. The highest value of the core tumor is written on the perimeter worst column. Area mean, area se, and area worst point at similar values related to the mean of the cancer cell areas, as described before. Smoothness mean is the mean for regional variations in radius range, smoothness se represents standard error of the mean of local variations in radius length, and the largest mean value is shown as smoothness worst. Compactness mean is a mean value of estimation of the perimeter and area, compactness se is used for standard error of compactness mean, and the highest mean value of the calculation is named compactness worst. Concavity mean shows the severity of concave portions of the shape, and concave points mean is the number of concave portions of the contour. Concavity se stands for the standard error of concave portions, while concave points se stands for

the standard error of the concave portions of the shape. Concavity worst and concave points worst stand for the highest mean value. Fractal dimension mean is the calculated mean value for coastline approximation, standard error of the coastline approximation is shown as fractal dimension se, and the highest mean value is fractal dimension worst [8], [15] [37].

3.9.2 Class attribute and distribution

The class attribute is diagnosis column. 37% data are malignant and 63% data are benign means 357 benign and 212 malignant from 569 instances.

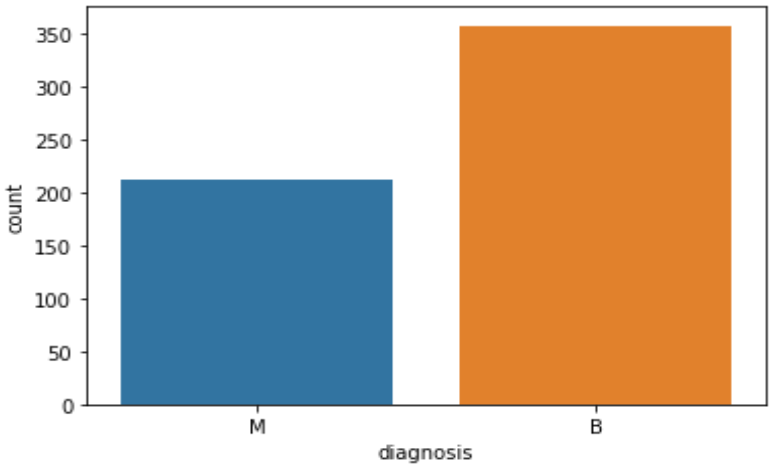

Figure 3.10: Counting diagnosis column.

3.9.3 Pair-plot

In the dataset, with diagnosis feature, there may some linear, non-linear relationship or cluster of data with one another are observed. We have only showed some pair-plot of attributes in the dataset.

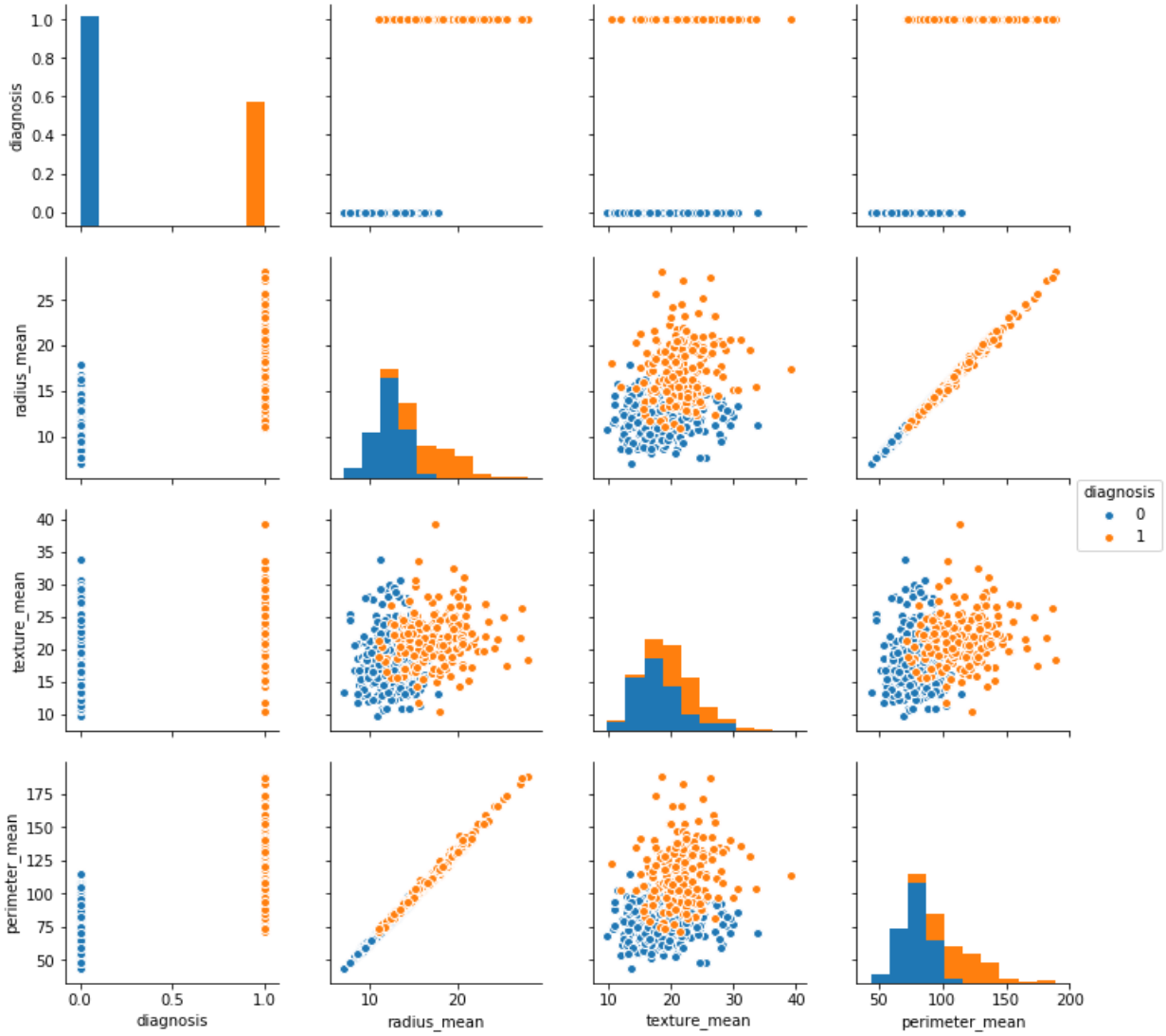

Figure 3.11: Observing the dataset by pair-plot

3.9.4 Correlation

It is a means of interpreting the interaction in the dataset between different variables and attributes. We may get some observations using Correlation, such as [38]:

- Some other attribute or a basis for another attribute relies on one or more attributes.
- For other attributes, one or more attributes are linked.

There are three types of correlation.

1. **Positive Correlation:** It means that if feature A increases then feature B also increases or if feature A decreases then feature B also decreases. Both features move

in tandem and they have a linear relationship [38]. Positive correlation values can be measured over 0.5 to 1.

2. **Negative Correlation:** It means that if feature A increases then feature B decreases and vice versa [38]. Negative correlation can be represented by -1.
3. **No Correlation:** It means there is no relationship in two variables [38]. It is measured in 0.

3.9.4.1 Importance

It is important for predicting the missing values from the data. It also gives us the knowledge about what models should be used.

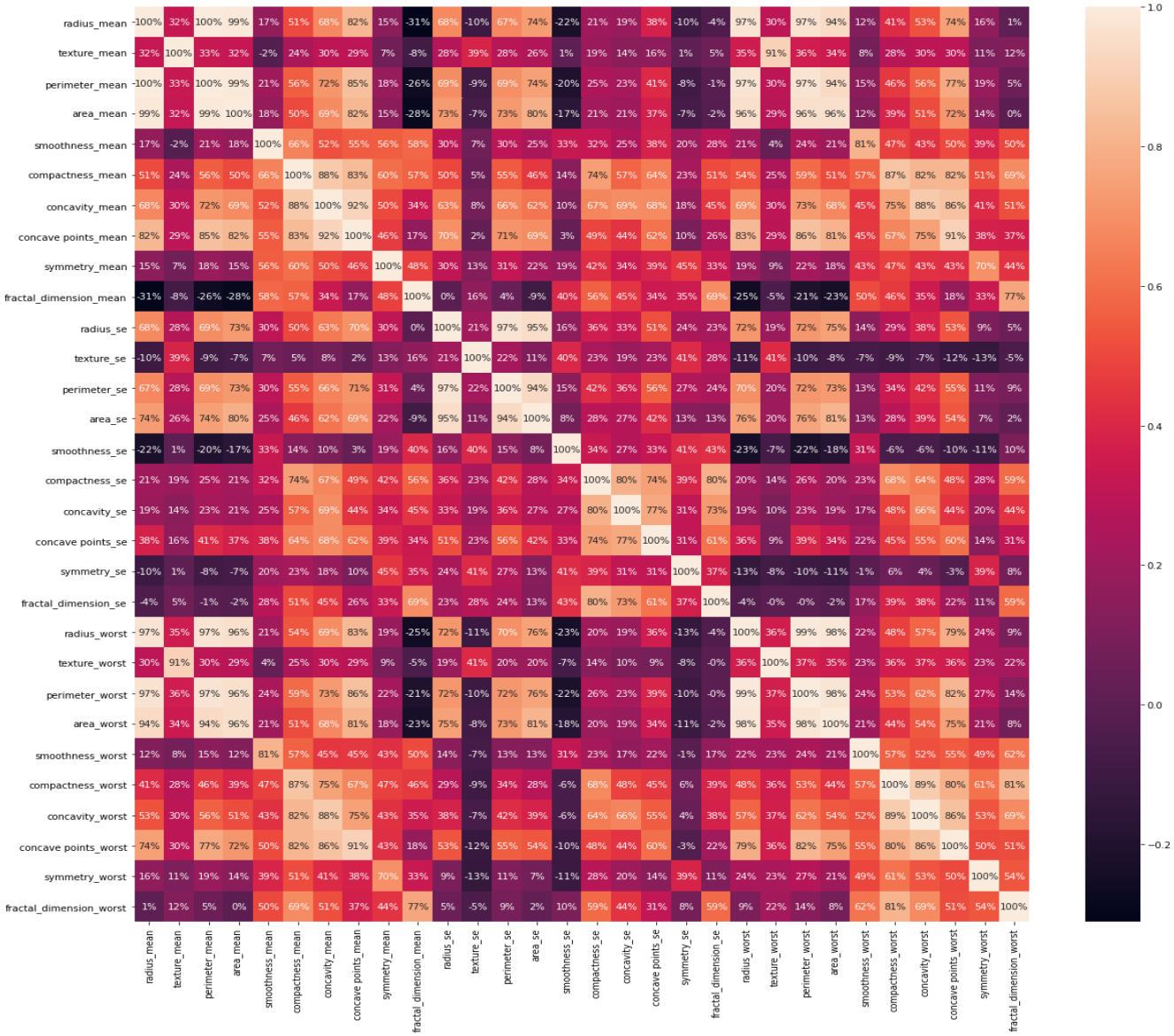

Figure 3.12: A mapping of correlation (in percentage) of WDBC dataset

It also gives us the idea about multi-collinearity. It means there are lots of independent values which can mislead the models of machine learning algorithms. But decision tree or boosted tree cannot be affected by multicollinearity. But Logistic Regression or Linear Regression can be affected by multicollinearity. By using PCA (Principal Component Analysis) or any dimensionality reduction method we can eliminate the multicollinearity in dataset. In the 3.12 Figure there are some values are in highly correlated. But multicollinearity effect can cause when training the dataset in the proposed models.

3.9.5 Conclusion

Correlation is very important to understand the dataset clearly. It also give the decision whether regression model should be used or not [39].

CHAPTER 4

Implementation of Methodologies

4.1 Introduction

This chapter plays a vital role in whole research paper. It is the heart of the total research. It covers the total data processing and implementations of machine learning algorithms for observing the accuracy measurement. We have given a snapshot of the methodology in previous chapter, at figure 3.9.

4.2 Data Preprocessing

Data preprocessing is one of the most data mining measures that deals with data planning and dataset transformation and aims to make the exploration of information more effective at the same time. Preprocessing requires many processes, such as washing, incorporation, conversion and elimination[40].

4.2.1 Preprocessing Techniques

Data preprocessing is one of the most data mining tasks which includes preparation and transformation of data into a suitable form to mining procedure. Data preprocessing aims to reduce the data size, find the relations between data, normalize data, remove outliers and extract features for data. It includes several techniques like data washing, incorporation, conversion and elimination [40]. Data preprocessing categories are shown in the below pictures.

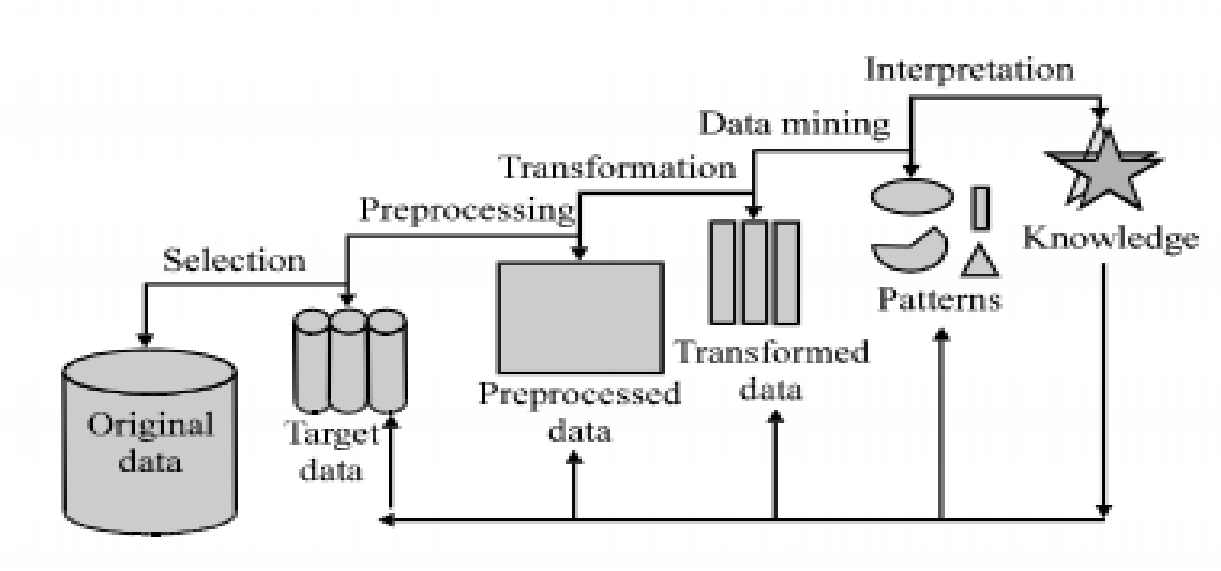

Figure 4.1 Knowledge Discovery Steps [40]

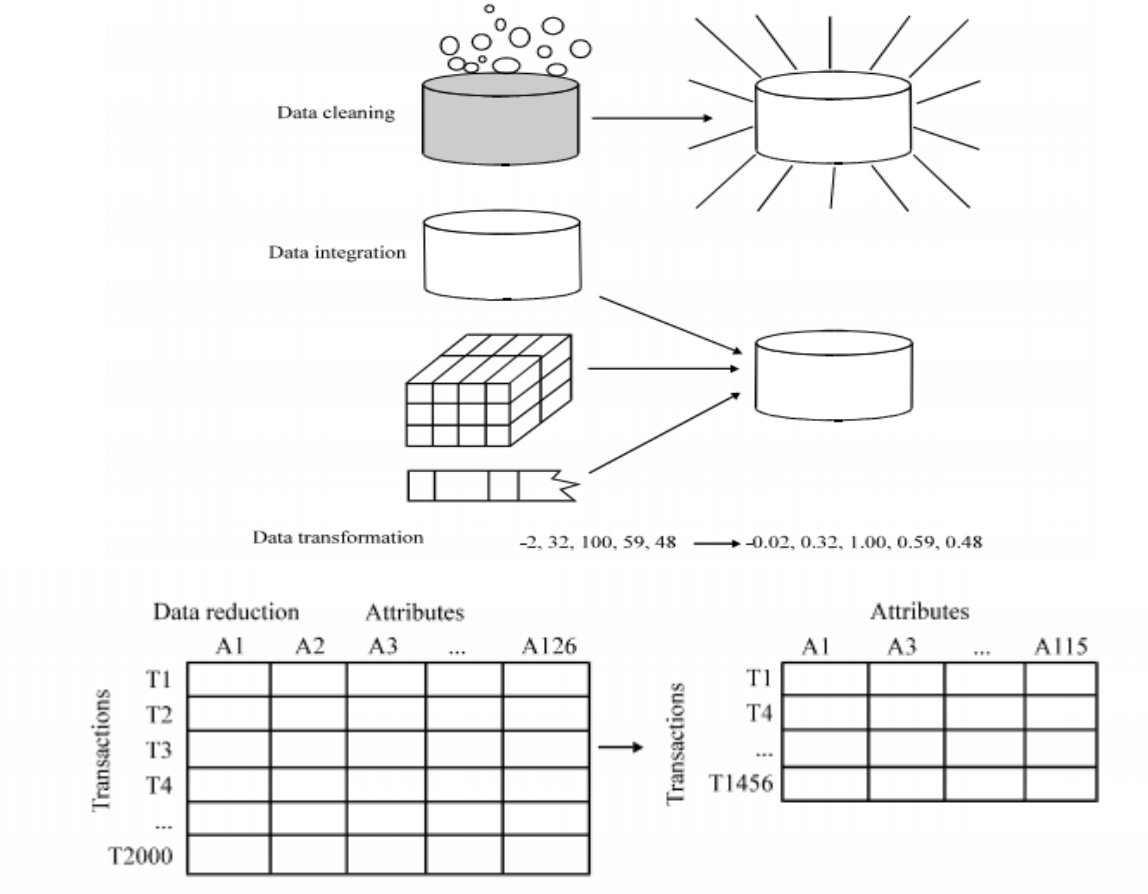

Figure 4.2: Preprocessing Forms [40]

4.2.1.1 Data Cleaning

Row data may have incomplete records, noise values, outliers and inconsistent data.

Data cleaning is a first stage of data preprocessing methods used to detect missed values, polished noise data, identify outliers and fix conflicting values. These dirty data will effects on mining procedure and led to unreliable and poor output [41]. Therefore, it is important for some data-cleaning routines to be used. Table 4.1 shows an example of dirty data.

Table 4.1: Examples of Dirty Data [40]

Dirty Data	Problems
Gender=S	Wrong Value
Address=0	Incomplete Record
C1_name =Rose M C2_name = R. Mohan	Duplicate Record
Name = Rose 15-10-2015	Multiple Values in single column

In our WDBC dataset, there is a column name “unnamed 32” which has no values or tuples. For this, we have to remove this column.

concavity_worst	concave points_worst	symmetry_worst	fractal_dimension_worst	Unnamed: 32
0.7119	0.2654	0.4601	0.11890	NaN
0.2416	0.1860	0.2750	0.08902	NaN
0.4504	0.2430	0.3613	0.08758	NaN
0.6869	0.2575	0.6638	0.17300	NaN
0.4000	0.1625	0.2364	0.07678	NaN
0.5355	0.1741	0.3985	0.12440	NaN
0.3784	0.1932	0.3063	0.08368	NaN

Figure 4.3: Removing missing values and column from WDBC dataset

4.2.1.2 Ignoring the Tuple

The choice is selected when the value of class label is not existing (it is used with classification mining task). This method is not effective but it is used when the tuple have several attributes with empty values [40]. Suppose the “id” of the patient can be ignored from WBCD dataset.

4.2.1.3 Filling the missing value manually

This approach in general requires human effort and time consuming. It cannot be used with the large size of dataset. But using the attribute mean to fill the missing value is largely used. This method works by replacing the missing value for that attribute. Using the most probable value to fill the missing value is also used [40].

	Variables			Missing values replaced by means		
Respondent	A	B	C	A	B	C
1	2	6		2	6	8
2		6	2	8	6	2
3		6		8	6	8
4	10	10	10	10	10	10
5	10	10	10	10	10	10
6	10	10	10	10	10	10
Average	*8*	*8*	*8*	*8*	*8*	*8*

Figure 4.4: Missing value handling [42].

4.2.1.4 Noise Data

One of the most problems which effects on mining process is noise. Noise is a random error or variance in a measured variable. Noise data means that there is an error in data or outliers which deviates from the normal. It can be corrected using binning, regression and clustering [40].

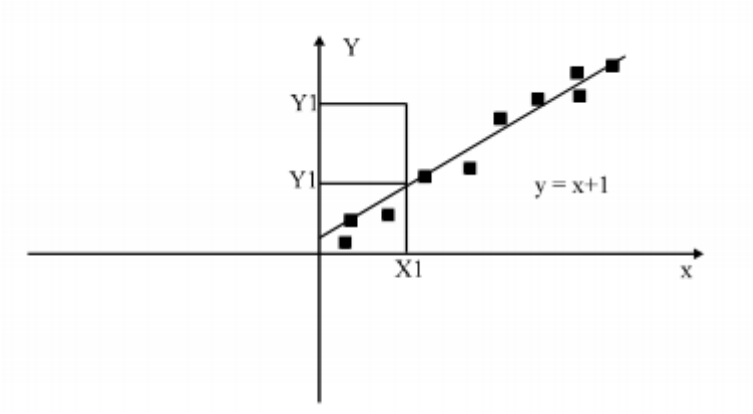

Figure 4.5: Regression

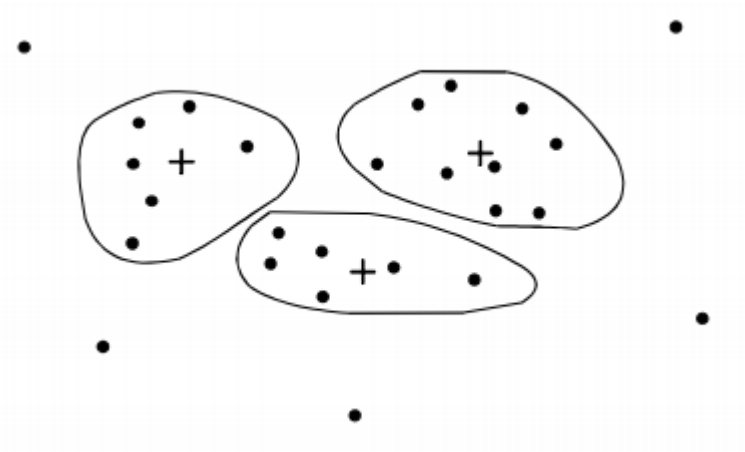

Figure 4.6: Clustering

4.2.1.5 Dealing with Categorical Data

If any data represents in qualitative format, then it is called categorical data. Some examples are in the following:

(i) Color conditions: “red”, “green”, “blue”, “yellow”, etc.

(ii) Tumor condition: “Benign” or “Malignant”

(iii) Exam grades: “A”, “B”, “C”, “D” and “F”

(iv) Gender: “Male”, “Female”, “Other” etc.

Encoding the data is very important because of some algorithms cannot handle categorical data. For this, the categorical data should be encoded into numerical data. Mapping is the key to encode the categorical data. Suppose in our WBCD dataset the class label is “B” and “M” means, Benign and Malignant. So we have to encode the data into 0 and 1.

B ⟶ 0

M ⟶ 1

```
Id      diagnosis

564     1
565     1
566     1
567     1
568     0
Name: diagnosis, dtype: int32
```

Figure 4.7: Encoding the class label data into 0 and 1.

4.2.1.6 Checking Data Types

Every columns of data should be checked. Some algorithms doesn't take different data types. In our WDBC dataset there are 32 columns. Every data types are shown in the below:

```
id                         int64
diagnosis                 object
radius_mean              float64
texture_mean             float64
perimeter_mean           float64
area_mean                float64
smoothness_mean          float64
compactness_mean         float64
concavity_mean           float64
concave points_mean      float64
symmetry_mean            float64
fractal_dimension_mean   float64
radius_se                float64
texture_se               float64
perimeter_se             float64
area_se                  float64
smoothness_se            float64
compactness_se           float64
concavity_se             float64
concave points_se        float64
symmetry_se              float64
fractal_dimension_se     float64
radius_worst             float64
texture_worst            float64
perimeter_worst          float64
area_worst               float64
smoothness_worst         float64
compactness_worst        float64
concavity_worst          float64
concave points_worst     float64
symmetry_worst           float64
fractal_dimension_worst  float64
dtype: object
```

Figure 4.8: A snapshot of WDBC data type

In the 4.9 figure every column are in float type of data except "diagnosis" column. So, "diagnosis" column should be converted into numerical values. As we early mentioned, categorical data can be converted into numerical data.

4.3 Training and Testing

After preparing the whole dataset, now we have to see how these machine learning algorithms works. That's why we have to train the model and by the test data we can see the

prediction result.

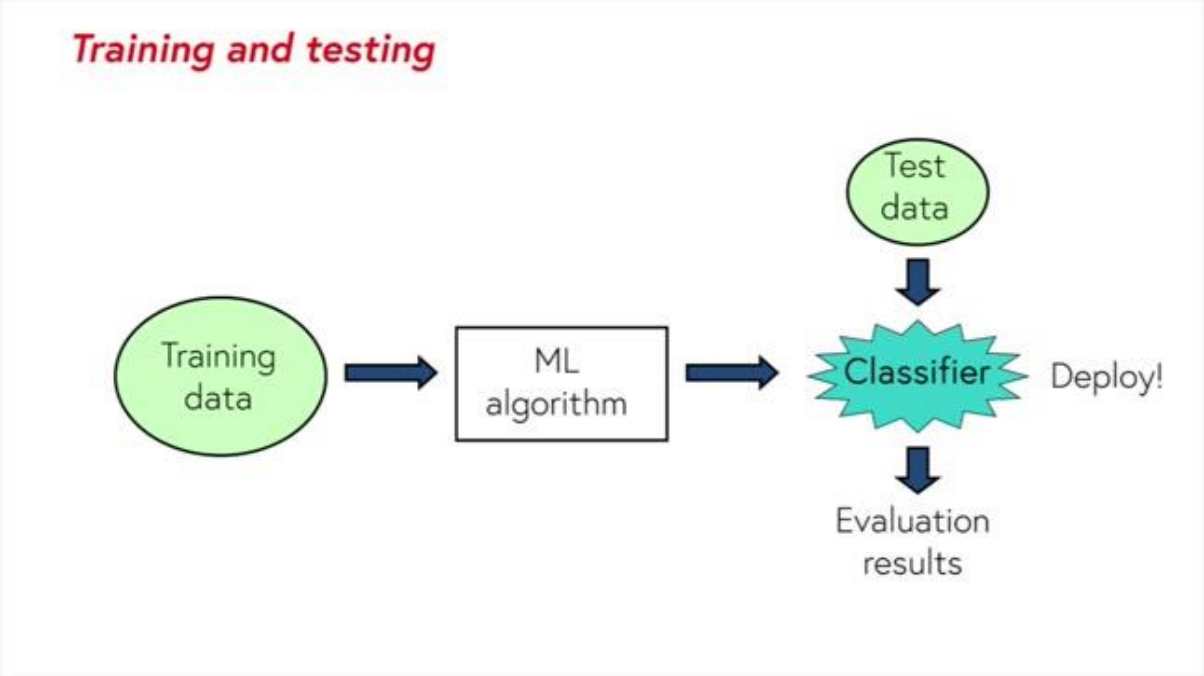

Figure 4.9: Training and testing of data[43]

To classify a dataset we have to train a model. But before training and testing we have to slice the dataset into dependent and independent data. In our WBCD dataset, “diagnosis” column is

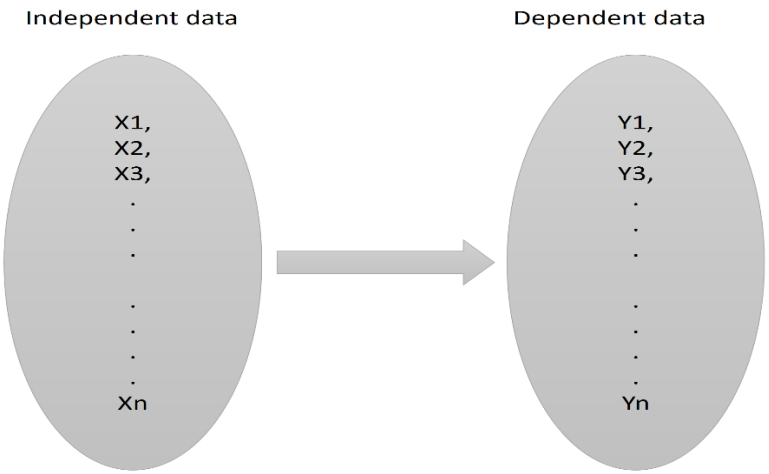

Figure 4.10: Independent and dependent dataset

dependent data and other columns are considered as independent data. Then we have to train and test the data.

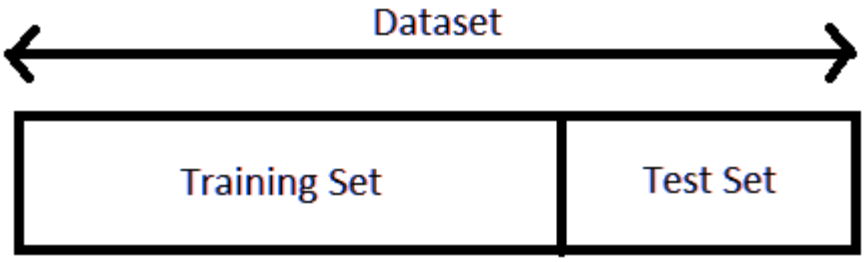

Figure 4.11: Training and testing dataset

Before observing the prediction, a machine learning model should be trained. Suppose we want to recognize a "cat" by some pictures. So a model has to be trained with various types of cat's pictures. After training the models now it has the ability to recognize the cat. By giving any picture of cat, it now give the decision of given new picture whether it will be a cat or not.
It seems like supervised learning. Our dataset is also a supervised learning dataset as it has the class label of "benign" and "malignant". Various types of sets are described in the following:

Training Set:

Here, we have the complete training dataset. We can extract features and train to fit a model and so on.

Validation Set:

Choosing the correct parameters for every estimator is critical. The training set can be split into a training set and a validation set. The model can be learned based on the validity test findings (for classifiers, changing parameters, instances). This will help us get the most optimized model. But some of them ignore the validation test as they consider in train set. Sometimes cross validation occurs for the validation set and training set. Cross validation means randomly split the training and validation data.

Testing Set:

Here, once the model is obtained, we can predict using the model obtained on the training set.

How much data should be trained and how much data should be tested is not specific. But is called that, at least 70% of data should be trained and rest of them are tested. The bigger the dataset to train is better. For our models we have trained the WBCD dataset by 4:1 proportion where 80% data are trained and 20% data are tested. The precision values of splitting training

and testing dataset based on the dataset size [44]. It now depends on us, what precision or accuracy we need to achieve based on our task.

4.4 Scaling the Training and Testing Data

It is actually a data preprocessing step. But after training and testing the independent data should be scaled. It is the method where the values are normalized. Suppose we have to scale a value in 0-5 0r 0 to 1 etc. It is very important because it measures the distance between data. If a distance is so large then the prediction will not be gain properly. For this we have to scale the data in a minimum distance or boundaries.

Let's have a look on gradient descent cost function. After scaling the minimum cost is found easily. Without it, optimization of cost function will be time consuming.

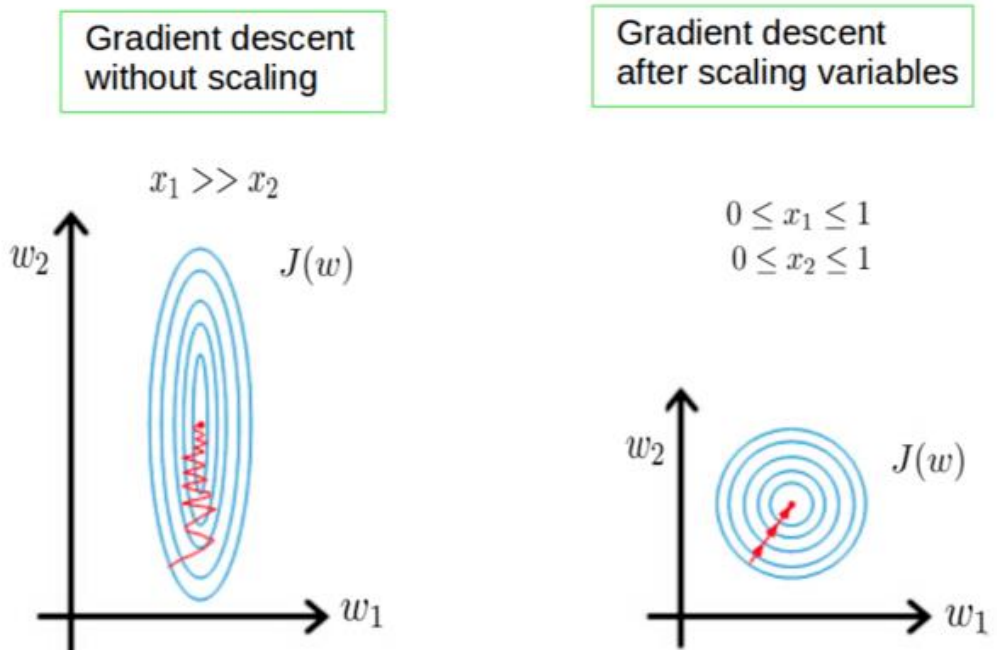

Figure 4.12: Scaling the value to optimize gradient descent[45]

4.5 Accuracy or Training and Testing Data

It is the main part of the whole chapter. In this section, we can observe the training data and testing data score.

4.5.1 Accuracy of Training Data

After scaling the accuracy of testing data is observed and they are recorded in the following:

Table 4.2 Accuracy of Training Data

Training accuracy of WDBC dataset			
(i)	Logistic Regression	0.98901	98.90%
(ii)	Decision Tree	1.0	100%
(iii)	Random Forest Classifier (number of estimator=60)	1.0	100%

4.5.2 Accuracy of Testing Data

For the accuracy of testing data, we have to use confusion matrix.

4.5.2.1 Confusion Matrix:

The efficiency of the classification model is defined by a confusion matrix. In other words, the matrix of confusion is a means of summarizing the success of classifiers. The following figure shows a basic representation of a confusion matrix:

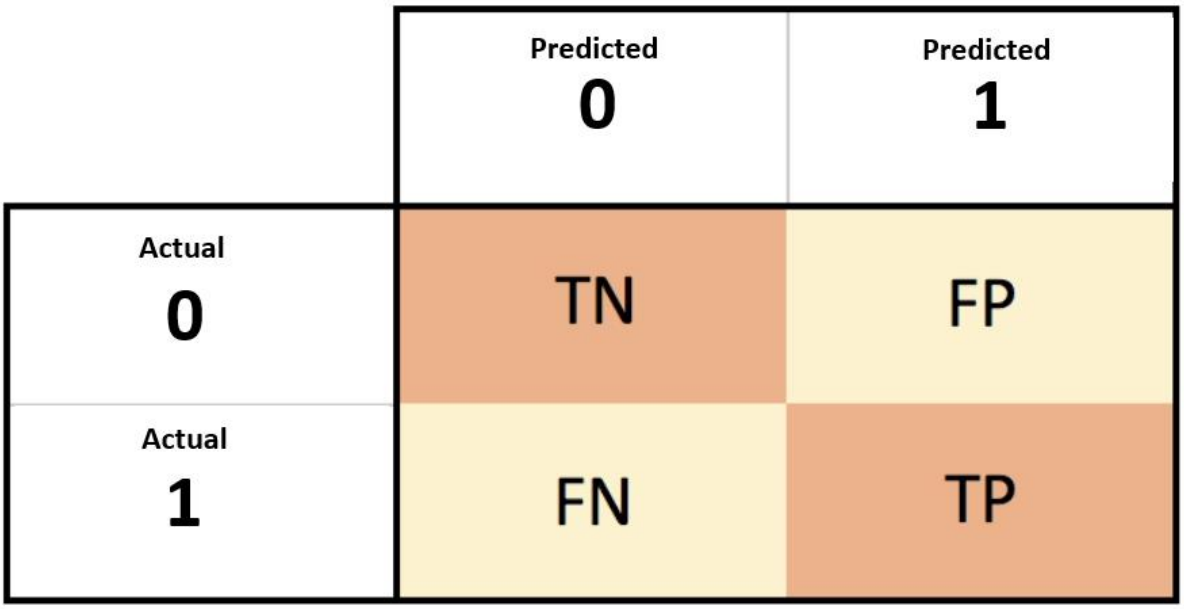

Figure 4.13: A confusion matrix

It has four part. They are:

(i) **True Positive:** This is when we have predicted a value is positive and actually true. Suppose the actual value is 1 and predicted value is also 1.

(ii) **True Negative:** This is when the actual value negative but true. Suppose the actual value is 0 and predicted value is also 0.

(iii) **False Positive:** When the actual value is false but predicted value is positive. Suppose the actual value is 0 but predicted value is 1.

(iv) **False Negative:** When the predicted value is negative and it's actually false. Suppose, we have predicted 0 but it's actually 1.

4.5.2.2 Confusion Matrix in WDBC Dataset

If we observe our dataset in confusion matrix, we can see that:

(i) Logistic Regression:

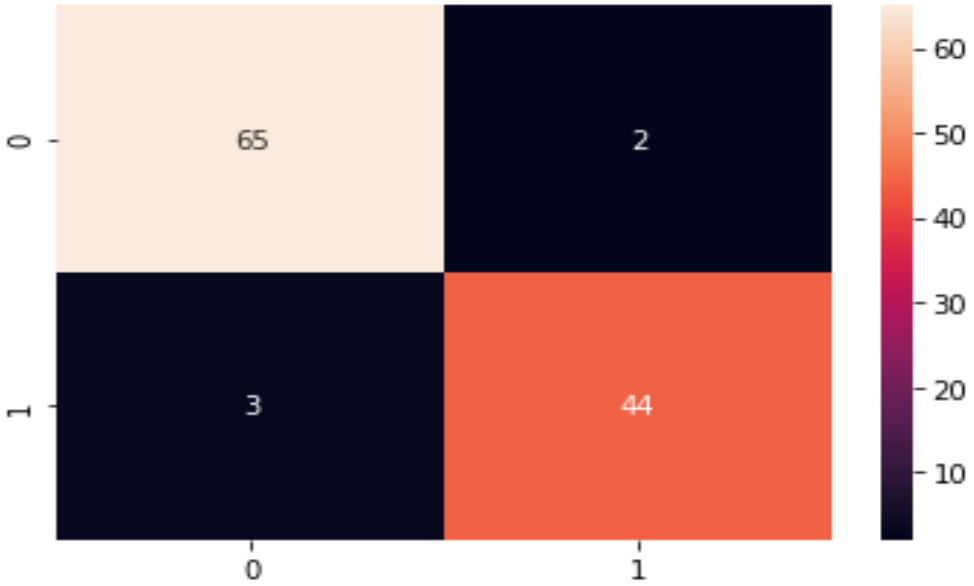

Figure 4.14: Confusion matrix of logistic regression.

(ii) Decision Tree:

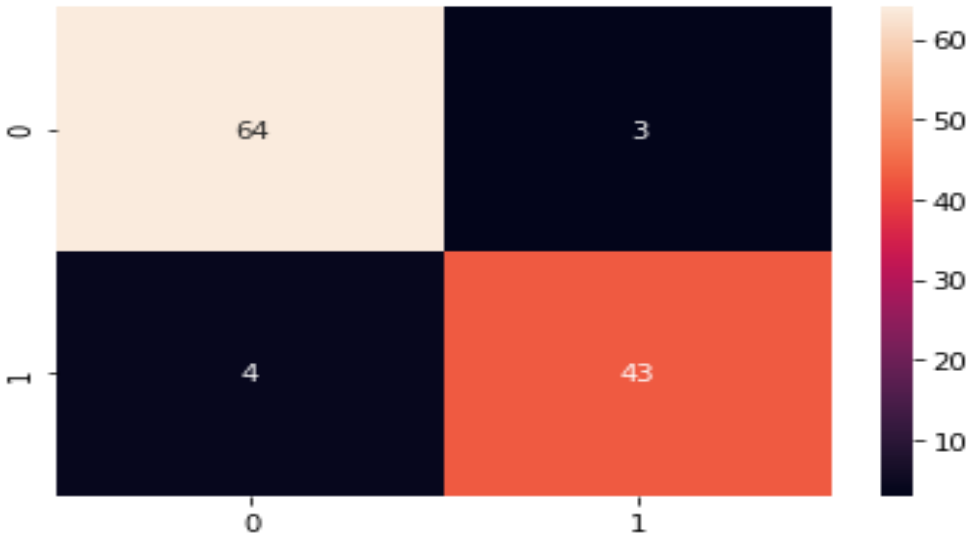

Figure 4.15: Confusion matrix of decision tree.

(iii) Random Forest Classifier (Number of Estimator=60):

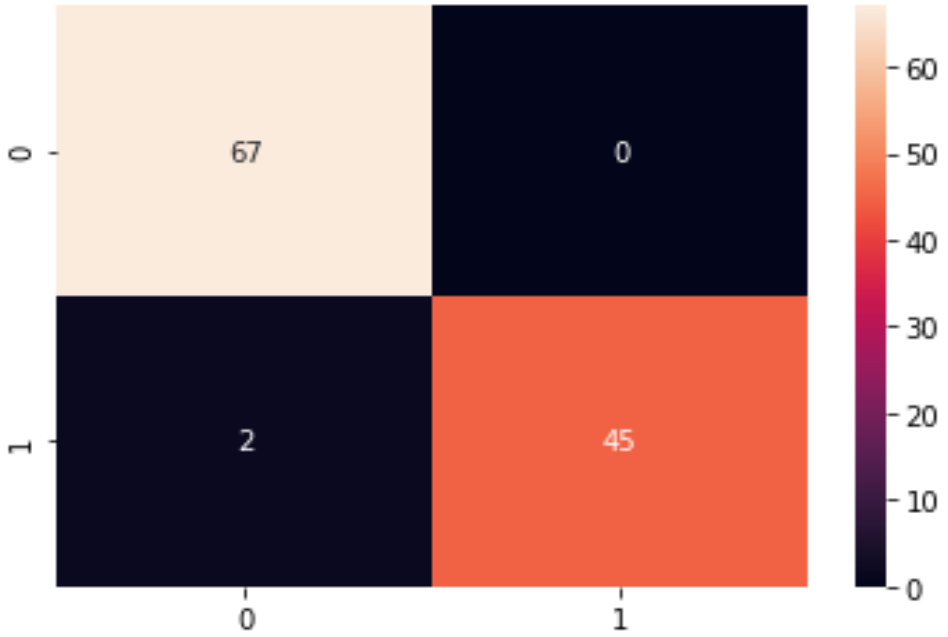

Figure 4.16: Confusion matrix of random forest.

If we consider confusion matrix with the words as True Negative [TN], False Positive [FP], False Negative [FN] and True Positive [TP], then the accuracy will be:

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

Testing accuracy of the models are shown in the following:

Table 4.3: Accuracy of Testing Data

Testing Accuracy of WDBC dataset			
(i)	Logistic Regression	0.95614	95.61%
(ii)	Decision Tree	0.93859	93.86%
(iii)	Random Forest Classifier (Number of Estimator =60)	0.98245	98.25%

4.6 Overfitting and Underfitting

Overfitting:

Overfitting occurs when the information and noise in the training data is learned by a model to the degree that it adversely influences the model's success on new data. This suggests that the noise or spontaneous variations in the training data are obtained by the algorithm and

taught as concepts. The concern is that these principles do not extend to new data and have a negative effect on the capacity of the models to generalize. Decision trees, for example, are a non-parametric machine learning method that is very versatile and susceptible to training data overfitting. This issue can be resolved by plucking a tree after learning to delete any of the information it has accumulated. Over all, it means decent training data response, bad generalization to other data.

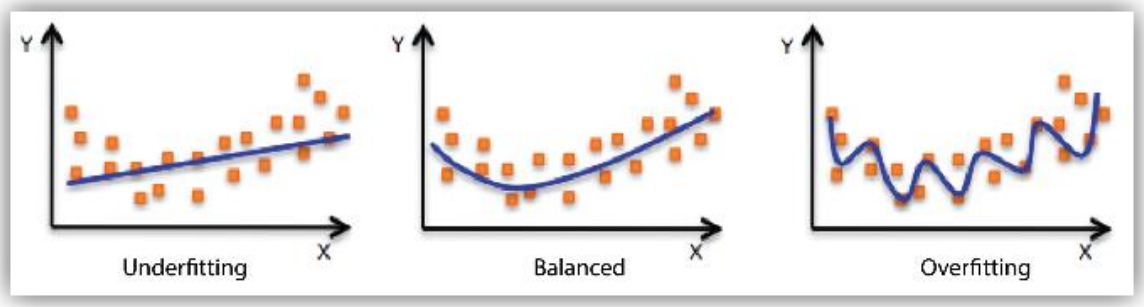

Figure 4.17: A snapshot of overfitting, balanced and underfitting data[46]

Balanced Data:

A balanced data set is a set that contains all elements observed in all time frame. Whereas unbalanced data is a set of data where certain years, the data category is not observed.

Underfitting:

Underfitting represents the idea which can neither model nor generalize the training data to new data. A weak machine learning model is not an acceptable model which will be noticeable as the training data will have bad outcomes. Over all, it provides other data with bad performances on the training data and weak generalization.

Let’s have a look to our training and testing accuracy.

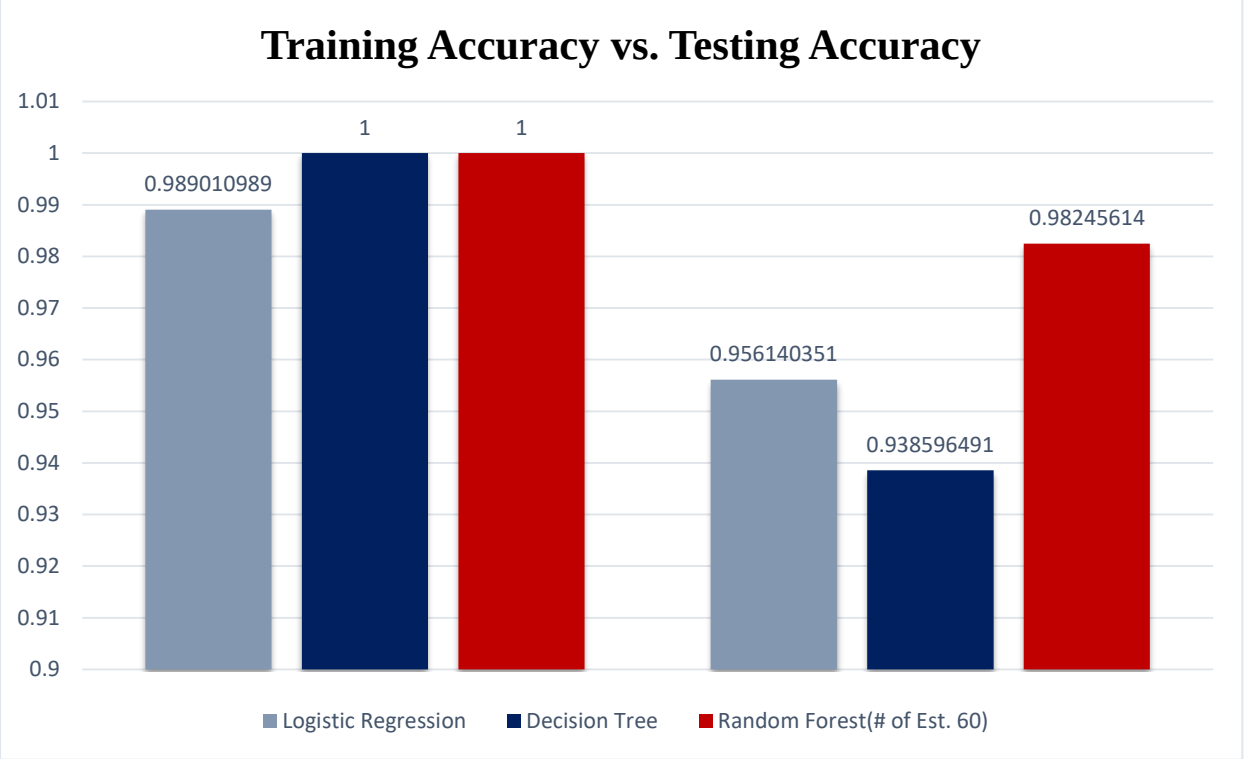

Figure 4.18: Training accuracy versus testing accuracy

On the 4.19 figure, we can see that the testing accuracy is slightly lower than the training accuracy. So, we have faced overfitting on this models. But underfitting is not happened to this dataset. So, on this point of view, dataset is tensed to balanced dataset and give better result.

4.7 Conclusion

On this chapter we have focused on the main theme of our research. After data preprocessing the dataset is fed into the models. By the confusion matrix, accuracy of every test dataset is observed. On this confusion matrix, it shows that random forest classifier gives the best accuracy. Then the training and testing accuracy is measured to see that our used models is overfit, underfit or balanced. From the training and testing score we can conclude that our dataset is showed slightly overfit.

CHAPTER 5

Result and Analysis

5.1 Introduction

In this chapter the result of this total thesis work is discussed. In section 5.2 the training and testing accuracy are analyzed. To find the best accuracy of Random Forest how the estimators are changed is discussed in section 5.3. Then the accuracy observation of the three used classifiers are discussed in rest of the chapter.

5.2 Result

In this chapter the result and the performance regarding the models are analyzed. We have trained the model and after testing the data different testing accuracy are showed in the following:

Table 5.1: Training and Testing Accuracy

Training Accuracy				
1. Logistic Regression			0.98901	98.90%
2. Decision Tree			1.0	100%
3. Random Forest	Number of estimator	30	1.0	100%
		45	1.0	100%
		60	1.0	100%
		80	1.0	100%
Testing Accuracy				
1. Logistic Regression			0.95614	95.61%
2. Decision Tree			0.93859	93.86%
3. Random Forest	Number of estimator	30	0.97368	97.37%
		45	0.96491	96.49%
		60	0.98246	98.25%
		80	0.96491	96.49%

In the above table 5.1, training and testing data showing little bit overfit in model as we mentioned in the previous chapter 4. But here random forest give the best solution in the above model.

5.3 Analysis

1. Random forest gives the better accuracy than Decision Tree. It's because of flexibility. Decision tree is just a single tree. But Random Forest is consisted by various trees. That's why Random Forest does better classification in dataset. By the increasing of number of trees, its classification accuracy will be getting better. In table 5.1, we can easily observe that, by increasing the number of estimators (trees), the accuracy is getting higher and lower. So we have to choose optimum number of trees. It is shown clearly in the following graph:

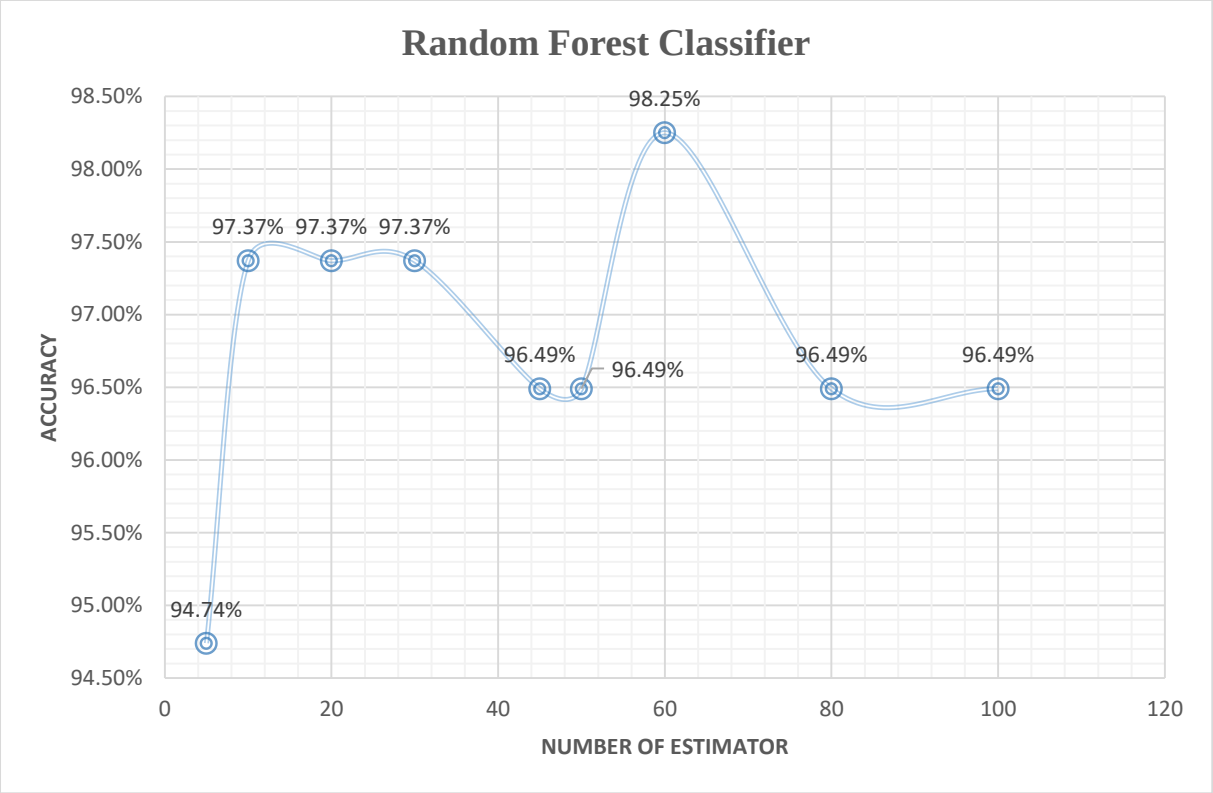

Figure 5.1: Random Forest Classifier Observation

In the figure 5.1, we can see that, the testing accuracy fluctuates on the increasing and decreasing of the estimators of Random Forest. We have found number of 60 estimators for higher accuracy value i.e. 98.25%. So choosing the optimum number of estimators is needed to find the proper accuracy. We have observed it within 100 estimators. Though all the estimators are in 100% training accuracy, the accuracy of testing are different.

2. For multi-collinearity in the dataset, Logistic Regression cannot perform better. For this it's accuracy is less than random forest. But it's accuracy is higher than Decision Tree.

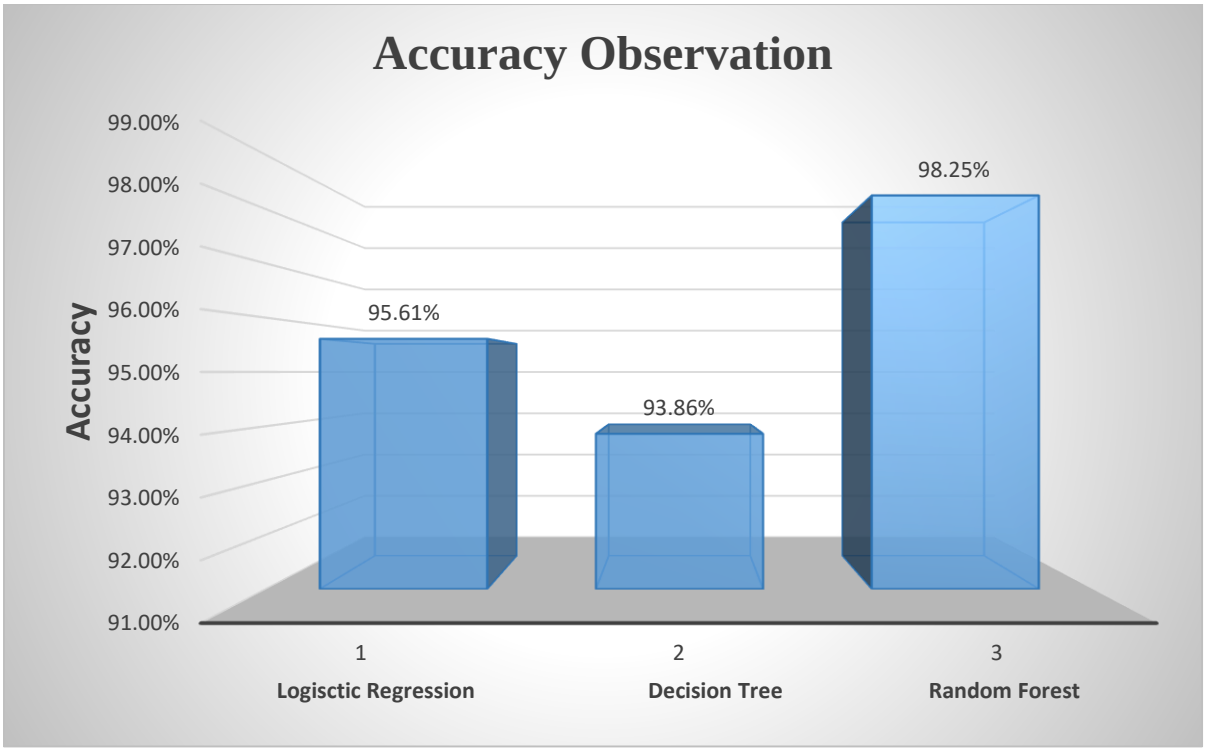

Figure 5.2: Accuracy observation

If we use both logistic regression and neural network modeling then it might lead to the best possible outcome prediction model [47].

3. Decision tree algorithm belongs to the family of supervised learning algorithms. The working principle of it is based on random selections. Positions of the features are selected randomly in the decision tree algorithm. Therefore, when the function runs several times, it is possible to get different accuracy results [8].

5.4 Comparison with Existing Work

In the referred research paper [8], the researcher used Logistic Regression, KNN, Support Vector Machine, Decision Tree, Naïve Bayesian, Random Forest and Rotation Forest for the detection of breast cancer. He divided the data set in three categories i.e. positively correlated, negatively correlated and all the features of dataset. He found the best accuracy on Logistic Regression which was 98.1% in all features included. But we have found Random forest as the best classifiers which gives 98.25% accuracy and Logistic Regression gives 95.61%. In his paper, Random forest gave 95.61%, 94.73% and 92.98% and Decision Tree gave 95.61%, 93.85, and 92.10% for all features, highly correlated features and low correlated features. The researcher also trained 80% of dataset and tested 20% of data. He used 50

estimators in Random Forest classifier. So, if we use all the features of dataset, Random Forest gives the best accuracy.

5.5 Conclusion

To detect the breast cancer Logistic Regression, Decision Tree and Random Forest classifiers are used. Among them Random Forest gives the best accuracy (98.25%). To find the accuracy, we have to find optimum number of trees i.e. 60. It gives better accuracy than existing work on this dataset.

CHAPTER 6

Conclusion and Future Works

6.1 Introduction

This chapter summarizes the whole thesis work, described in last five chapters. Section 6.2 describes the total summary of the thesis work, section 6.2 and 6.3 sort out the limitations and future works and finally 6.4 concludes the thesis work.

6.2 Summary

Data science is a multidisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data. The primary subtopics of computer science include analytics, data processing, data visualization, machine learning, deep learning, and artificial intelligence. While data science was established in the 1990s, nowadays the relevance of this area is acknowledged. In numerous reports, it is stated that the volume of data in the world is increasingly growing, and more than half of the overall amount of data still accounted for the unstructured data form. Therefore, data science has become an essential issue in any field to make data understandable [8].

The goal of this research paper was to predict the breast cancer dataset whether it is benign tumor or malignant tumor. First of all, we have to observe the dataset removing missing attribute column, checking missing values, encoding categorical values to numerical values, scaling the independent data, visualization of the dataset to see the correlation and then training and testing the dataset. Then they are learned in Logistic Regression, Decision Tree and Random Forest with number of trees. From these models, Random Forest gives the higher accuracy that is 98.25% in 60 estimators (trees). It’s because of dealing with multi-correlated value is stronger than other algorithms. On the other hand, Random Forest is more flexible to classify any dataset. Random Forest is derived from many decision trees. For this, it’s very flexible to classify the dataset more accurately. Logistic regression cannot deal with multi-correlated or highly correlated dataset.

Attributions to the cause of death in those with breast cancer may depend on numerous reasons related to the specifications of the patient. As a consequence of breast cancer, any particular cause will decrease the chance of death. Nevertheless, these observations underline the importance of early diagnoses encountered by patients with a history of breast cancer, both current and past. Our analysis confirms the significance of highly successful early detection of women with breast cancer[8].

6.3 Limitations

In this research work, there are several limitations. Some are mentioned in the following:

(i) In data preprocessing, smaller value of data are not checked. These smaller data can't show effectiveness in mathematical calculations. For this, these smaller data should be normalized.

(ii) In train-test split of the dataset, we have only split the dataset by 80% of training data and 20% of test data. But cross validation method is not used. By using this cross validation, it may reduce the overfitting of the models.

6.4 Future Works

We have done a classification of breast cancer dataset with only three algorithms. Our main goal is to early detection of breast cancer. As it is very important to detect the cancer some of the future work are described in the below:

1. Except this dataset, more cancer dataset can be observed in the way of this methodology. To observe the similarity or dissimilarity, it is very important to train and test more dataset.
2. For the highly correlated value, the highly correlated value can be categorized individually. After this, the prediction can be measured.
3. Dimensionality reduction techniques (Principal Component Analysis, High Correlation Filer, Low Variance Filer etc.) can be used to see the accuracy may varies or not.
4. Except train-test split method, cross validation can be used.

6.5 Conclusion

In this thesis works, the prediction of the breast cancer is observed by using Logistic Regression, Decision Tree and Random Forest. After data preprocessing, data visualization these classifiers are used and they give their individual performance. Though this thesis works has some limitations, it contains a lot of information of classifiers and implementation strategies of predicting breast cancer. By approaching this mentioned future works one can extend this thesis work.

REFERENCES

[1] J. Calle, 'Breast cancer facts and figures 2003–2004', *Am. Cancer Soc.*, pp. 1–27, 2004.

[2] M. Karabatak and M. C. Ince, 'An expert system for detection of breast cancer based on association rules and neural network', *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3465–3469, 2009.

[3] 'World Cancer Report 2008.'https://www.cabdirect.org/cabdirect/abstract/20103010665 (accessed Dec. 21, 2020).

[4] M. Mori *et al.*, 'Diagnostic accuracy of contrast-enhanced spectral mammography in comparison to conventional full-field digital mammography in a population of women with dense breasts', *Breast Cancer*, vol. 24, no. 1, pp. 104–110, 2017.

[5] Md. M. Hasan, Md. R. Haque, and M. Md. J. Kabir, 'Breast Cancer Diagnosis Models Using PCA and Different Neural Network Architectures', in *2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)*, Rajshahi, Bangladesh, Jul. 2019, pp. 1–4, doi: 10.1109/IC4ME247184.2019.9036627.

[6] B. A. U. and S. Freedman, 'Breast Cancer.', presented at the The 2nd edition,Springer Science and Business Media, 2008.

[7] A. M. Elsayad, 'Diagnosis of Breast Cancer using Decision Tree Models and SVM', *Int. J. Comput. Appl.*, vol. 83, no. 5, p. 12.

[8] M. F. Ak, 'A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications', *Healthcare*, vol. 8, no. 2, p. 111, Apr. 2020, doi: 10.3390/healthcare8020111.

[9] K. Juneja and C. Rana, 'An improved weighted decision tree approach for breast cancer prediction', 2020.

[10] W. Yue, Z. Wang, H. Chen, A. Payne, and X. Liu, 'Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis', *Designs*, vol. 2, no. 2, Art. no. 2, Jun. 2018, doi: 10.3390/designs2020013.

[11] S. Aruna, S. P. Rajagopalan, and L. V. Nandakishore, 'Knowledge based analysis of various statistical tools in detecting breast cancer', *Comput. Sci. Inf. Technol.*, vol. 2, no. 2011, pp. 37–45, 2011.

[12] V. Chaurasia and S. Pal, 'Data mining techniques: to predict and resolve breast cancer survivability', *Int. J. Comput. Sci. Mob. Comput. IJCSMC*, vol. 3, no. 1, pp. 10–22, 2014.

[13] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, ‘Using machine learning algorithms for breast cancer risk prediction and diagnosis’, *Procedia Comput. Sci.*, vol. 83, pp. 1064–1069, 2016.

[14] H. Wang and S. W. Yoon, ‘Breast cancer prediction using data mining method’, in *IIE Annual Conference. Proceedings*, 2015, p. 818.

[15] O. L. Mangasarian, W. N. Street, and W. H. Wolberg, ‘Breast cancer diagnosis and prognosis via linear programming’, *Oper. Res.*, vol. 43, no. 4, pp. 570–577, 1995.

[16] W. William H. and W. N. Street, ‘UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set’. https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic) (accessed Dec. 19, 2020).

[17] ‘Biopsy’, *CIRSE*. https://www.cirse.org/patients/ir-procedures/biopsy/ (accessed Feb. 19, 2021).

[18] ‘Diagnostic Mammography During the COVID-19 Pandemic | Mayfair Diagnostics’, Apr. 22, 2020. https://www.radiology.ca/article/diagnostic-mammography-during-covid-19-pandemic (accessed Feb. 19, 2021).

[19] Atul, ‘AI vs Machine Learning vs Deep Learning’, *Edureka*, Jun. 08, 2018. https://www.edureka.co/blog/ai-vs-machine-learning-vs-deep-learning/ (accessed Feb. 19, 2021).

[20] N. Donges, ‘A complete guide to the random forest algorithm’, *Built In*. https://builtin.com/data-science/random-forest-algorithm (accessed Dec. 21, 2020).

[21] J. Jordan, ‘Hyperparameter tuning for machine learning models.’, *Jeremy Jordan*, Nov. 02, 2017. https://www.jeremyjordan.me/hyperparameter-tuning/ (accessed Dec. 21, 2020).

[22] Md. M. Islam, Md. R. Haque, H. Iqbal, Md. M. Hasan, M. Hasan, and M. N. Kabir, ‘Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques’, *SN Comput. Sci.*, vol. 1, no. 5, p. 290, Sep. 2020, doi: 10.1007/s42979-020-00305-w.

[23] R. Nisbet, J. Elder, and G. Miner, *Handbook of statistical analysis and data mining applications*. Academic Press, 2009.

[24] H. A. Elsalamony and A. M. Elsayad, ‘Bank direct marketing based on neural network and C5. 0 Models’, *Int J Eng Adv Technol IJEAT*, vol. 2, no. 6, 2013.

[25] C.-M. Chao, Y.-W. Yu, B.-W. Cheng, and Y.-L. Kuo, ‘Construction the model on the breast cancer survival analysis use support vector machine, logistic regression and decision tree’, *J. Med. Syst.*, vol. 38, no. 10, p. 106, 2014.

[26] A. Chakure, ‘Decision Tree Classification’, *Medium*, Nov. 06, 2020. https://medium.com/swlh/decision-tree-classification-de64fc4d5aac (accessed Dec. 21, 2020).

[27] Y.-Q. Liu, C. Wang, and L. Zhang, ‘Decision Tree Based Predictive Models for Breast Cancer Survivability on Imbalanced Data’, in *2009 3rd International Conference on Bioinformatics and Biomedical Engineering*, Beijing, Jun. 2009, pp. 1–4, doi: 10.1109/ICBBE.2009.5162571.

[28] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, ‘An introduction to decision tree modeling’, *J. Chemom.*, vol. 18, no. 6, pp. 275–285, Jun. 2004, doi: 10.1002/cem.873.

[29] L. Rokach, *DECISION TREES*.

[30] L. Fahrmeir, T. Kneib, S. Lang, and B. Marx, *Regression*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013.

[31] S. Polamuri, ‘How the logistic regression model works’, *Dataaspirant*, Mar. 02, 2017. https://dataaspirant.com/how-logistic-regression-model-works/ (accessed Dec. 21, 2020).

[32] P. Sarkar, ‘Machine Learning: What is Logistic Regression?’, Sep. 23, 2019. https://www.knowledgehut.com/blog/data-science/logistic-regression-for-machine-learning (accessed Dec. 21, 2020).

[33] A. Worster, J. Fan, and A. Ismaila, ‘Understanding linear and logistic regression analyses’, *CJEM*, vol. 9, no. 02, pp. 111–113, Mar. 2007, doi: 10.1017/S1481803500014883.

[34] S. P. Morgan and J. D. Teachman, ‘Logistic Regression: Description, Examples, and Comparisons’, *J. Marriage Fam.*, vol. 50, no. 4, pp. 929–936, 1988.

[35] D. Dua and C. Graff, ‘UCI machine learning repository, 2017’, *URL Httparchive Ics Uci Eduml*, vol. 37, 2019.

[36] ‘Breast Cancer Wisconsin (Diagnostic) Data Set’. https://kaggle.com/uciml/breast-cancer-wisconsin-data (accessed Dec. 21, 2020).

[37] E. Aličković and A. Subasi, ‘Breast cancer diagnosis using GA feature selection and Rotation Forest’, *Neural Comput. Appl.*, vol. 28, no. 4, pp. 753–763, 2017.

[38] W. Badr, ‘Why Feature Correlation Matters A Lot!’, *Medium*, Jan. 22, 2019. https://towardsdatascience.com/why-feature-correlation-matters-a-lot-847e8ba439c4 (accessed Dec. 21, 2020).

[39] L. Zhao, Y. Chen, and D. W. Schaffner, ‘Comparison of Logistic Regression and Linear Regression in Modeling Percentage Data’, *Appl. Environ. Microbiol.*, vol. 67, no. 5, pp. 2129–2135, May 2001, doi: 10.1128/AEM.67.5.2129-2135.2001.

[40] S. A. Alasadi and W. S. Bhaya, ‘Review of Data Preprocessing Techniques in Data Mining’. Journal of Engineering and Applied Science, 2017, [Online]. Available: https://www.researchgate.net/profile/Suad_Alasadi/publication/320161439_Review_of_Data_Preprocessing_Techniques_in_Data_Mining/links/59d143d64585150177f3d15b/Review-of-Data-Preprocessing-Techniques-in-Data-Mining.pdf.

[41] Maini and M.N., ‘Survey on data preprocessing concept applicable in data mining: Mining.’ 2013.

[42] T. Bock, ‘5 Ways to Deal with Missing Data in Cluster Analysis | Displayr’. https://www.displayr.com/5-ways-deal-missing-data-cluster-analysis/ (accessed Dec. 21, 2020).

[43] ‘Training and testing’, *FutureLearn*. /info/blog (accessed Feb. 19, 2021).

[44] D. Padmanabhan, S. Bhat, S. Shevade, and Y. Narahari, ‘Topic Model Based Multi-Label Classification from the Crowd’, *ArXiv Prepr. ArXiv160400783*, 2016.

[45] aijayanta Roy, ‘All about Feature Scaling. Scale data for better performance of… | by Baijayanta Roy | Towards Data Science’, Apr. 06, 2020. https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35 (accessed Feb. 19, 2021).

[46] ‘Model Fit: Underfitting vs. Overfitting - Amazon Machine Learning’. https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html (accessed Feb. 19, 2021).

[47] J. V. Tu, ‘Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes’, *J. Clin. Epidemiol.*, vol. 49, no. 11, pp. 1225–1231, Nov. 1996, doi: 10.1016/S0895-4356(96)00002-9.