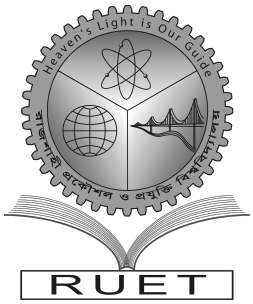
Heaven's Light is Our Guide



DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Rajshahi University of Engineering & Technology, Bangladesh

Coronary Artery Disease(CAD) Prediction Using Machine Learning Methods

Author

Dibbo Barua Chamak

Roll No. 1603117

Department of Computer Science & Engineering

Rajshahi University of Engineering & Technology

Supervised by

Dr. Md. Ali Hossain

Professor

Department of Computer Science & Engineering

Rajshahi University of Engineering & Technology

ACKNOWLEDGEMENT

First of all, I want to express my gratitude to the Almighty God for providing me with the chance and motivation to finish our thesis work.

I want to sincerely thank my supervisor and show him my gratitude and respect. Professor of computer science and engineering at Rajshahi University of Engineering and Technology, Rajshahi, Dr. Md. Ali Hossain. He has not only provided me with the technical instructions, guidance, and papers I've needed to finish the work; he has also continuously encouraged me and offered support whenever he saw fit. His constant assistance was the most effective instrument I had at my disposal to get the job done. He was always available to me at any moment of the day when I was caught in any complicated issues or circumstances. Without his genuine concern, this work would not have taken the current form that it does.

I also want to express my gratitude to all of the instructors at Rajshahi University of Engineering and Technology's Computer Science and Engineering departments for their time-to-time helpful advice and inspiration.

Finally, I would want to express my gratitude to my parents, friends, and well-wishers for their ongoing inspiration and numerous supportive contributions to this work.

October 30, 2022 Dibbo Barua Chamak

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 ***"Coronary Artery Disease(CAD) Prediction Using Machine Learning Methods"*** *submitted by* ***Dibbo Barua Chamak, Roll:1603117*** *in partial fulfillment of the requirement for the award of the degree of Bachelor of Science in Department of 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. Md. Ali Hossain
Professor
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi-6204

External Examiner

_____

Rizoan Toufiq
Assistant Professor
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi-6204

ABSTRACT

Cardiovascular disease (CVD) is the main cause of death worldwide.17.9 million individuals died from CVD in the world in 2019, accounting for 32% of all deaths.Heart attacks and strokes were the main causes of 85% of these deaths.It appears that Coronary Artery Disease is one of the common cardiovascular defect.For a long time, researchers have been interested in coronary artery disease (CAD).Large numbers of lives could be saved if CAD disease is detected early.A number of researchers made significant contributions to the classification of CAD disease prior to attaining this goal, but in each case, either the precision was insufficient or the number of features used was excessive. The problem of dimensionality affects classification when the number of features is huge.The objective of this thesis is to assess models and select one that can correctly predict CAD.In this thesis, four classifiers were used to categorize CAD, and their performances were assessed in terms of accuracy, precision, recall, and f1-score. These classifiers were Naive Bayes, Random Forest, Support Vector Machine, and Neural Network(Multilayer Perceptron).The random forest classifier produces 93.55% on the heart statlog Cleveland Hungary final dataset as GridSearchCV was used to find the best suitable hyperparameter that results in a better accuracy.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CVD - Cardiovascular Disease

CAD - Coronary Artery Disease

ECG - Electrocardiogram

SVM - Support Vector Machine

RF - Random Forest

NB - Naive Bayes

NN - Neural Network

EDA - Exploratory Data Analysis

Chapter 1

Introduction

On a Coronary Artery disease dataset, this thesis investigated the predictions of Random Forest, Support Vector Machine, Naive Bayes, and Neural Network classifiers.To assess the effectiveness of the classifiers, traditional machine learning approaches were used.This chapter describes the work's challenges, motivations, objectives, and thesis organization.

1.1 Introduction

Cardiovascular disorders are among the prevalent illnesses in both industrialized and developing nations, and they are widely acknowledged as the leading cause of death worldwide[8].In 2019, 17.9 million people died from CVDs, accounting for 32% of all global deaths.85% of these deaths were caused by a heart attack or a stroke. [9].Coronary vascular diseases can be caused by any illness or condition that affects the heart or its veins[10], or the blood circulation system (CVDs)[11].The most frequent Cardiovascular abnormality appears to be Coronary Artery Disease (CAD)[9]Coronary artery disease (CAD), also termed as coronary heart disease, is a leading risk factor for death[12].It is the most common form of cardiovascular problem, with numerous subtypes including stable and unstable angina, myocardial infarction, and sudden cardiac death[13]. Chest pain or discomfort is a typical symptom that might move to the shoulder, arm, spine, neck, or maxilla[14]. It may arise as heartburn with usual symptoms after exercise or restless stress on occasion, last for a few minutes, and resolve with comfort or rest[14].Breathing difficulties may arise, and there may be no symptoms at times; nevertheless, in many situations, the initial indicator is a heart seizure, heart malfunction, or an abnormal heartbeat[15]. Smoking, diabetes, high blood pressure and cholesterol, lack of activity, poor nutrition, depression,

and excessive alcohol consumption are all risk factors for CAD[15][16][17].

According to the latest WHO data published in 2020- ”Coronary Heart Disease Deaths in Bangladesh reached 108,528 or 15.16% of total deaths”[18]. Bangladesh ranks 118 in the world with an age-adjusted death rate of 94.27 per 100,000 inhabitants[18].

Previously, various researchers made significant advances to diagnosing CAD with greater accuracy by focusing on recognition with fewer classes.In this thesis, the classifiers considered are Naive Bayes(NB), Support Vector Machine(SVM), Neural Network(NN) Random Forest(RF).Using these four classifers, we tried to show our result tried to make a comparison among the classifiers.

1.2 Challenges

Coronary artery disease is the leading cause of death worldwide and a major public health concern[19].There are a few challenges that were experienced for the diagnosis of CAD disease in this thesis.One of the key challenges in accurately detecting CAD is selecting the important features.One of the most difficult challenges is discovering and cleaning large datasets.Choosing the models that will be best suited for detecting CAD from all of the existing models is a major difficulty.

1.3 Motivation

From the above discussion, it is exhibited that with millions of people dying from CAD each year, early predictions can prolong from heart disease problems. Using machine learning methods can help detecting CAD in earlier stage with higher accuracy .Thus it can play huge impact on the medical system. So, this actually motivates me to work on this topic to predict CAD in early stage.

1.4 Problem Statement

Cardiovascular diseases are among the most common ailments in both developed and developing countries, and they are commonly regarded as the main cause of mortality globally[8].In both low and middle-income countries, more than one-third of cardiovascular disease deaths

occur.As a result, this is one of the most catastrophic crises confronting the globe today.People affected by cardiovascular diseases or at high risk, particularly those experiencing signs of hypertension, hyperlipidaemia, or who already have any established disorders, require early detection and appropriate counseling to save their lives from a potentially fatal disease.
South Asians are also at risk.South Asians account for one-fourth of the global population.This region has the highest birth rate in the world.People in South Asia, particularly in India, Bangladesh, Pakistan, and Sri Lanka, are particularly vulnerable.Their way of living is the primary cause of their terrible heart condition.Despite having a large population, they account for 60% of the world's heart disease sufferers. As a result, the current high number of patients suffering from coronary heart disease has become a serious public health concern.Coronary artery disease has become a more serious problem in Bangladesh than any other condition. Because of an unhealthy lifestyle, the majority of people are at high risk.People in Bangladesh live less healthy lives than people in other countries throughout the world.They do not willingly obey the set of rules of medical sciences.In Bangladesh, cardiovascular disease kills 2.56 lakh people, with noncommunicable diseases accounting for nearly 30% of all deaths.In comparison to one or two years, the death rate grows rapidly.Experts have classified Bangladesh as a "red zone" for this potentially fatal heart ailment[20] .According to the most recent WHO data published in 2020, 108,528 people died from coronary artery disease in Bangladesh, accounting for 15.16 percent of all deaths[18]. Bangladesh ranks 118 in the world with an age-adjusted death rate of 94.27 per 100,000 inhabitants[18]. The diagnosis and treatment of the CAD disease have become an extremely urgent work to do. Its increasing in alarming rate date rate is also increasing . So I want to work in this topic. It contains analysis of the CAD disease and detecting it using four classifiers.

1.5 Research Objectives

Predicting CAD had become very popular method now-a-days. For finding the better accuracy, classifiers selection is one the most important task.

The objective of our thesis is give below:

- To predict Coronary Artery Disease with higher accuracy level .
- To predict CAD using less feature than the previous works .

- To show the Comparison between multiple classifiers.
- To compare the performance of my work and the previous works

1.6 Research Contribution

In this thesis, machine learning classifiers were used to predict Coronary Artery Disease. In our study, we first displayed the patients' dataset and then realized the patients' CAD situations. Four classifier predictions were displayed and compared. To improve the findings, some hyperparameter tuning was given to each classifier, and the best classifier was chosen from among the classifiers.

1.7 Organization of Thesis

The remainder of the thesis work is structured as follows: :

Chapter 2 - Coronary Artery Disease

This chapter is dedicated about defining Coronary Artery Disease, causes of CAD, precautions of CAD survival rate of CAD.

Chapter 3 - Background Study and Literature Review

This chapter is dedicated for background study and literature review. Different types of statistical and machine learning approaches that have been proposed earlier will be discussed here.

Chapter 4 - Research Dataset

Here, we will discuss about the dataset that used for this research.The dataset attributes will be discussed here.

Chapter 5 - Implementation

Here, we will discuss the experimental setup and the whole implementation process for the prediction of CAD.

Chapter 6 - Results and Performance Analysis

The overall findings of our proposed methodology will be discussed in this chapter, and their performances will be compared to other state-of-the-art methodologies.

Chapter 7 - Conclusion and Future Works

Finally,in this chapter we will discuss about our final outlines. These outlines are conclusion, few limitations and our future work.

1.8 Conclusion

Coronary Artery Condition is a disease that must be discovered early and precisely in order to be treated. The thesis proposes an automated system that can solve the specified challenge.

Chapter 2

Background Study & Literature Review

2.1 Introduction

This chapter discusses several cardiovascular disease scenarios.It also provides a clear glimpse of the coronary artery vessels and aids in determining how it operates, its symptoms and its risk factor.This chapter focuses on the variables that cause disorder.This chapter also goes into machine learning.

2.2 Coronary Artery Disease(CAD)

Coronary artery disease (CAD), widely known as coronary heart disease, occurs when the coronary arteries become narrowed or cholesterol plaques develop on the arterial walls.The coronary arteries are the blood vessels that supply oxygen and blood to the heart[1].

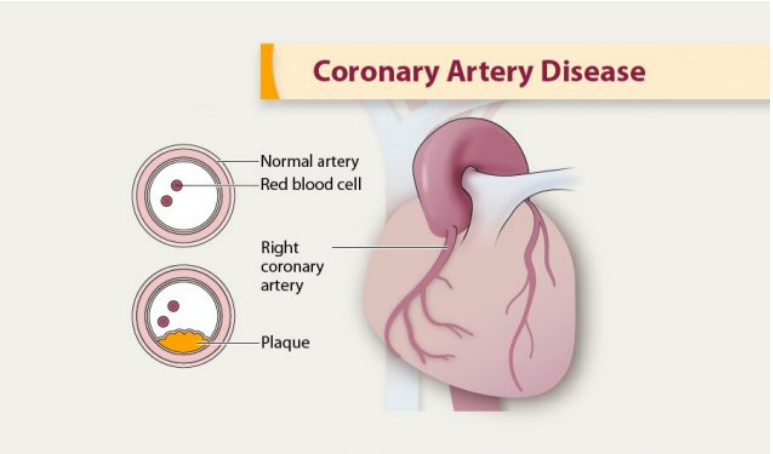


Figure 2.1: Coronary Artery Disease[1]

Cholesterol accumulates on the artery walls, forming plaques, which is a common cause of coronary artery disease. These plaques can narrow the arteries, reducing blood flow to the heart, or they can cause irritation and stiffening of the blood vessel walls.A clot can occasionally obstruct blood flow, causing serious health problems.The network of oxygen-supplying blood capillaries on the surface of the heart is made up of coronary arteries.If these arteries become constricted, especially during severe exercise, the heart may not get enough oxygen-rich blood.A heart attack can occur as a result of CAD. [21].

2.2.1 What Causes CAD?

Plaque, a substance composed of cholesterol, fat, and other substances, begins to form on the inside walls of your blood vessels as early as childhood. It accumulates over time. This causes the hardening and constriction of arteries, which doctors refer to as ”atherosclerosis.”In rare circumstances, plaque might crack or burst.Platelets will try to repair the artery by forming a blood clot in response. This accumulation impedes the passage of blood through the arteries, similar to clogs in a drainpipe.The heart is supplied with oxygen and nutrients through blood.If you don’t receive enough, it might cause chest discomfort and shortness of breath (angina).Without sufficient oxygen, the heart might weaken.This can cause a heartbeat irregularity (arrhythmia).It can also cause heart failure, which occurs when the heart is unable to pump enough blood to meet the body’s needs. A heart attack may occur if a plaque develops so big that it obstructs blood flow to the heart muscle.Most heart attacks, however, are caused by the rupture of smaller plaques[22].

2.2.2 What Are the Signs and Symptoms?

You may not have any symptoms in the early stages.If the plaque builds up and inhibits blood flow to the heart muscle, you may suffer shortness of breath or weariness, especially during vigorous exertion. The most prevalent sign of coronary artery disease is angina, or chest discomfort. Some may confuse it for acid reflux or indigestion. With angina, chest discomfort is present. The ache might also be felt in the shoulders, arms, back, or jaw. One cam feel:

- Tightness
- Discomfort

- Pressure
- Heaviness
- Squeezing
- Burning
- Aching
- Numbness
- Fullness[22].

Chest pain Particularly in the middle or left side of the chest, and lasting a few minutes or reoccurring.There may be a sensation of pressure, squeezing, fullness, or discomfort. Some may confuse it with indigestion or heartburn.

Discomfort in any part of upper body. Discomfort might also be felt in the shoulders, arms, back, or jaw.

Shortness of breath with or without chest discomfort

Nausea or vomiting with lightheadedness, dizziness, or cold sweat

Symptoms of a heart attack in women typically differ from males. While chest pain is still the most prevalent warning sign, Shortness of breath, extreme tiredness, nausea, vomiting, and back or jaw ache are more common in women[22].

2.2.3 Risk Factor

Coronary artery disease (CAD) is more common as you age or if you have a family history of it. However, you may control several additional risk factors, such as:

- High cholesterol and triglycerides
- High cholesterol and triglycerides
- High blood pressure
- Smoking

- Syndrome of Metabolic Syndrome
- Diabetes
- Obesity and being overweight
- Lack of exercise
- Stress, depression, and anger
- Unhealthy diet
- Excessive alcohol consumption
- Obstructive sleep apnea [22].

2.2.4 Diagnosis

Your doctor will examine you and discuss your symptoms, hazards, and family medical history. You may also be given the following tests:

Electrocardiogram (ECG or EKG) This analyzes the electrical activity of the heart and can evaluate heart disease

Stress test Typically, this comprises walking on a treadmill or riding a stationary bike at a doctor's office while your EKG, pulse rate, and hypertension are monitored.

Chest X-ray

Cardiac catheterization A technique in which a doctor inserts a catheter (a thin, flexible tube) into a blood vessel in the arm or leg and guides it to the heart.The doctor injects a contrast agent through the catheter and then uses X-ray images to view the inside of the heart. [22].

2.3 Machine Learning

Machine learning is an Artificial Intelligence branch that enables computers to learn and develop without even being explicitly programmed[23].Machine learning may specialize computer programs that can obtain data and learn on their own. Observations like examples, knowledge, or

direction are the building blocks of learning; they enable us to look for patterns in data and refine our judgments over time.Machine learning's goal is to give computers the ability to figure things out for themselves, without human intervention, and to act accordingly.

There are several accessible algorithms for machine learning. Algorithms that utilize machine learning to discover patterns in enormous amounts of data utilize statistics. And data in this context refers to a multitude of things, such as numbers, words, images, clicks, and anything else comes to mind. Data is likely to be supplied to a machine-learning algorithm if it is routinely processed digitally.There are four primary categories of machine learning algorithms. There are three types of learning: supervised, unsupervised, reinforced, and semi-supervised. The following is a summary of these three machine learning approaches.

2.3.1 Supervised Learning

Using labeled examples, supervised machine learning algorithms apply prior knowledge to new data to predict upcoming scenarios. By examining a known training dataset, the learning algorithm generates a predicted output function.After adequate training, the system can generate objectives for every new input. Additionally, it may compare its output to the proper, intended output to identify faults and adapt the model accordingly[24].

2.3.2 Unsupervised Learning

The fundamental idea of supervised learning is to seek for situations when the supervision signal is regarded as a target value or name. This form of signal does not exist in unsupervised learning.As a result, we'd rather go shooting without oversight or supervision.This simply means that we are completely on our own and must figure out what is going on.We are not, however, entirely in the dark. Everyday, we engage in this solitary study. We don't have labels for data points in unsupervised learning, but we do have the data points itself. This indicates that references will be derived from observations inside the supplied data.

This learning approach is used by self-organizing neural networks to find hidden patterns in unlabeled input data.Unsupervised is the ability to seek and arrange information without the use of an error signal to assess the prospective response. [25].

2.3.3 Reinforcement Learning

Rather than specifying learning techniques, reinforcement learning is distinguished by identifying a learning issue[26].A reinforcement learning strategy is any approach to problem-solving that is well-balanced. Reinforcement learning differs from supervised learning, which is the sort of learning studied in the majority of contemporary machine learning, statistical pattern recognition, and artificial neural network science. When we cannot reach optimality, we must experimentally evaluate the performance of RL algorithms[27].The fact that reinforcement learning explicitly takes into account the whole problem of goal-directed agents The ability to communicate with a dangerous environment is also an important aspect[26].A reinforcement learning strategy is any balanced way for solving the issue. Reinforcement learning differs from supervised learning, which is the type of learning investigated in the most contemporary machine learning, statistical pattern recognition, and artificial neural network science. When optimality cannot be reached, RL algorithm performance must be experimentally evaluated[27].Reinforcement learning is further distinguished by its comprehensive consideration of a goal-directed agent's interaction with a potentially hazardous environment.Instead of learning methods, reinforcement learning is defined by a learning issue.

2.3.4 Semi-Supervised Learning

Semi-supervised learning, also known as partly supervised learning, is a kind of machine learning that builds classifiers using both supervised and unsupervised[28].Using semi-supervised learning, one may deal with such datasets without having to make the trade-offs that supervised or unsupervised learning require. This framework is controlled by semi-supervised learning algorithms.

1. This method employs a tiny set of labeled sample data to educate itself, resulting in a model with limited training.

2. The partly trained model identifies the unlabeled data. Because the sample-labeled data gathering has a lot of serious issues.

3. Labeled and pseudo-labeled datasets are integrated to create a single method that encompasses the descriptive and predictive components of supervised and unsupervised learning.

2.4 Literature Review

As Coronary Artery Disease has become a major worry in the modern world, a number of research are now being conducted. Experts are seeking to forecast CAD illness and identify characteristics among local patients.

- Purba, Fredrick Dermawan and Hunfeld, Joke AM and Fitriana, Titi Sahidah and Iskandarsyah, Aulia and Sadarjoen, Sawitri S and Busschbach, Jan JV and Passchier, Ja[29] have shown comparison between two models - SVM and ANN model. SVM shown higher accuracy and better performance than the ANN model. they have used multilayer perceptron with three hidden layer in ANN model. This paper suggested to use the other Data mining algorithms to improve the positive predictive value of the disease prediction.

- In their study, Z. Liu, Shulong Zhang, and colleagues[30] proved that heart failure risk may be predicted using sequential EHR data modeling. The most significant contribution of their article is that they sought to forecast heart failure using a neural network to predict the risk of cardiac illness using electronic medical data.

- Hlaudi Mosima Daniel Masethe and Anna Masethe and colleagues[31] suggested an approach for predicting cardiac disease using classification algorithms.

- In their article, they employed Waikato Environment for knowledge analysis, and this was used to estimate its skill at discovering, analyzing, and forecasting trends. They collected information from a clinic in South Africa. They employed the J48, REPTREE, Naive Bayes, and BAYES S NET classifiers to predict the results.

- Ashok Kumar Dwivedi et al.[32] evaluated the performance evaluation of several machine learning approaches for the prediction of cardiac disease. In his article, he incorporated the methods ANN, KNN, Support Vector Machine, Logical Regression, Classification Tree, and Naive Bayes.

- Data mining was utilized by Mustafa Badshah, Vishwa Bhagwat, and colleagues[33] to forecast cardiac disease. They utilized a Cleveland Clinic Foundation-collected UCI dataset. They utilized PCI for data preprocessing prior to utilizing Decision Tree, Random Forest, Naive Bayes, Support Vector Machine, MLP Classifier, and Logical Regression techniques to obtain accurate prediction.

2.5 Conclusion

This chapter is the most important portion of this study. It outlines the backdrop of the research. In the first section of this chapter, coronary artery disease is defined, along with its causes, symptoms, and risk factors. Various types of machine learning were also emphasized.

Chapter 3

Research Dataset

The dataset that was used for this study will be explained in this chapter. This chapter will discuss the dataset's background analysis as well as some intriguing data-related findings.

3.1 Dataset Description

A dataset named "Heart Disease Dataset (Comprehensive)[34]" is used for this research and is found from kaggle.This dataset was compiled by fusing many datasets that were previously available separately but had never been integrated. The result is the largest heart disease dataset available for research because they aggregated them into 11 common features. Dataset of instances:

- Cleveland: 303
- Hungarian: 294
- Switzerland: 123
- Long Beach VA: 200
- Stalog (Heart) Data Set: 270[34]

total 1190 instances.

This dataset has a target variable and 11 characteristics. There are 6 nominal and 5 numerical variables in it. The following is a full description of every feature:

1. Age: Patients' ages are expressed in years (Numeric).

2. Sex: Patient gender (Male - 1, Female - 0) (Nominal).

3. Chest Pain Type: Type of chest pain experienced by patient categorized into-

1. typical
2. typical angina
3. non- anginal pain
4. asymptomatic (Nominal)

4. resting bp s: Blood pressure in mm/HG in resting mode (Numerical).

5. cholestrol: Serum cholestrol in mg/dl (Numeric)

6. fasting blood sugar: Blood sugar levels after fasting > 120 mg/dl are represented as

1 in the event of true and

0 in the case of false (Nominal) [35]

7. resting ecg: The ECG result when at rest is represented by three separate numbers.

0 : Normal

1: Abnormality in ST-T wave

2: Left ventricular hypertrophy (Nominal)

8. max heart rate: Maximum heart rate achieved (Numeric)

9. exercise angina: Angina induced by exercise

0 depicting NO

1 depicting Yes (Nominal)

10. oldpeak: In compared to the resting state, exercise caused ST-depression (Numeric).

11. ST slope: Peak workout ST segment assessed in terms of slope

0: Normal

1: Upsloping

2: Flat

3: Downsloping (Nominal)[35]

NO	age	sex	chest pain type	cholesterol	fasting blood sugar	resting ecg
0	40	1	2	140	289	0
1	49	0	3	160	180	0
2	40	1	2	130	283	1
3	37	0	4	138	214	0
4	48	1	3	150	195	0

Table 3.1: Dataset Outlook

NO	max heart rate	exercise angina	oldpeak	ST slope	target
0	172	0	0.0	1	0
1	156	0	1.0	2	1
2	98	0	0.0	1	0
3	108	1	1.5	2	1
4	122	0	0.0	1	0

Table 3.2: Dataset Outlook

3.2 Exploratory Data Analysis (EDA)

Exploratory data analysis (EDA) is a technique used in statistics that involves the use of statistical graphics and other data visualization methods to explore data sets in order to highlight their significant properties. In contrast to traditional hypothesis testing, EDA explores what can be learned from the data outside of a strictly modeled context. Possible uses of a statistical model are discussed. Since 1970, John Tukey has advocated for exploratory data analysis to encourage statisticians to delve deeper into the data and perhaps generate new ideas that can inform future data collection and experiments. Initial data analysis (IDA) focuses on completing the data by adding or replacing values, adjusting variables as necessary, and verifying assumptions for model fitting and hypothesis testing; EDA goes beyond this. [36].

3.2.1 Distribution of CAD disease (target variable)

In the Datset, 53% instances have CAD disease 47% instances are normal.The dataset is balanced having 629 CAD disease patients and 561 normal patients.

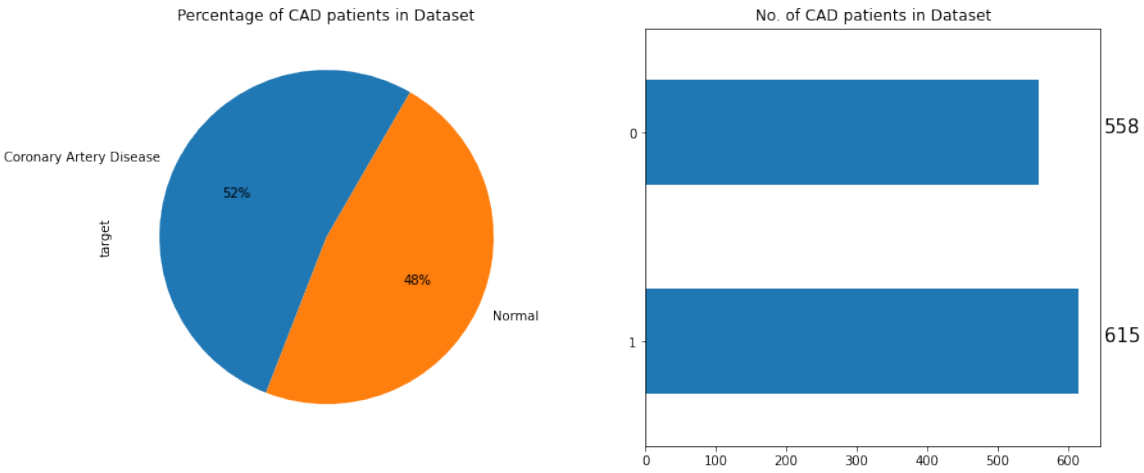


Figure 3.1: The frequency of CAD disease(target variable)

In the above figure, 0 means Normal patients and 1 means CAD patients.

3.2.2 Checking Gender & Agewise Distribution

Males outnumber females in this dataset, despite the fact that the average age of patients is about 55.

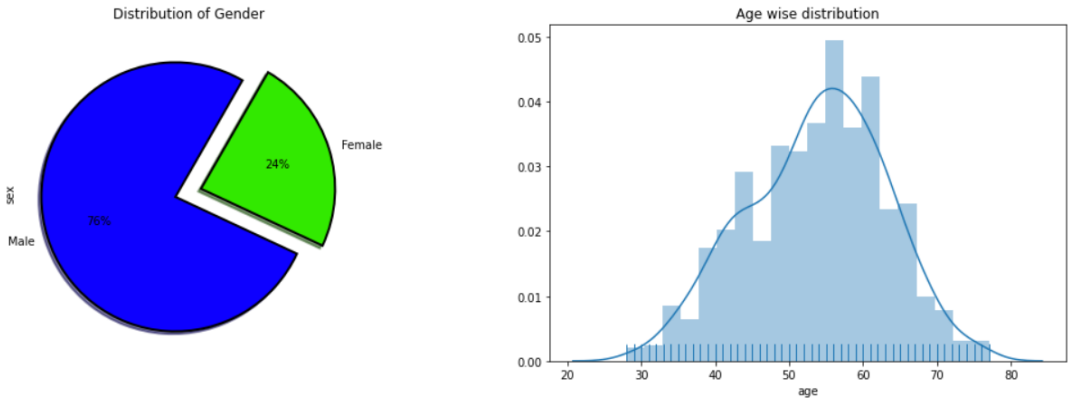


Figure 3.2: Gender & Agewise Distribution

3.2.3 Distribution of Chest Pain Type

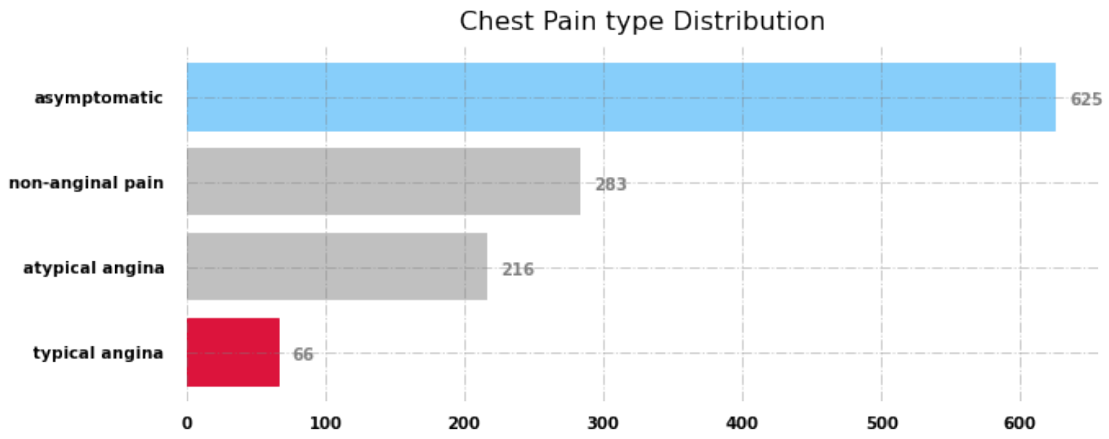


Figure 3.3: Chest pain type distribution

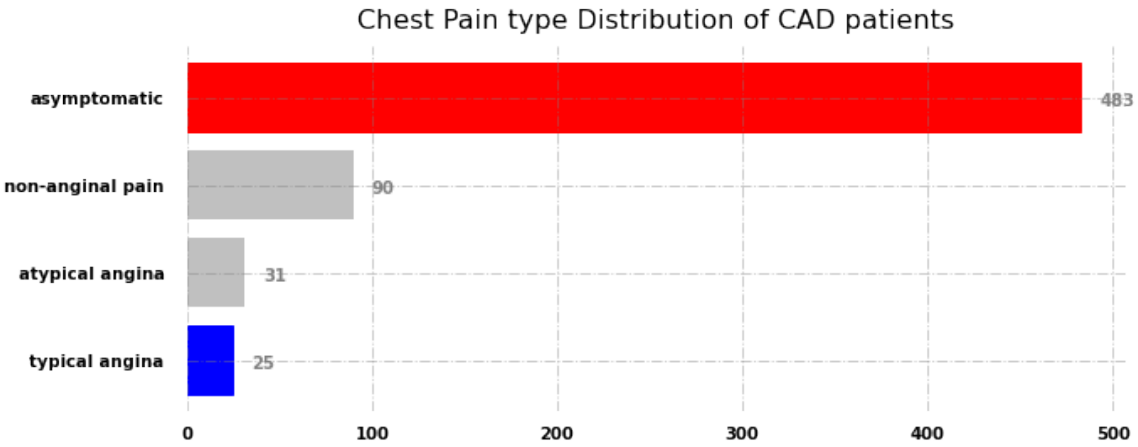


Figure 3.4: Chest pain type distribution of CAD patients

As we can see from the accompanying plot, 76% of heart disease patients who experience chest pain do so without any other symptoms. asymptomatic heart attacks, also referred to as silent myocardial infarction (SMI), account for 45–50% of cardiac-related morbidities each year as well as some premature deaths. Males are more prone than females to develop SMI in middle life, with male cases being twice as common. SMI is known as a silent killer since the symptoms are so weak compared to a real heart attack. When compared to the symptoms of a typical heart attack, which include severe chest pain, stabbing pain in the arms, neck, and jaw, sudden shortness of breath, perspiration, and dizziness, symptoms of SMI are far more fleeting and frequently mistaken for everyday discomfort[37].

Target	0	1
chest_pain_type		
asymptomatic	25.31	76.79
atypical angina	32.98	4.93
non-anginal pain	34.4	14.31
typical angina	7.31	3.97

Table 3.3: target Vs Chest Pain type

3.2.4 Distribution of Rest ECG

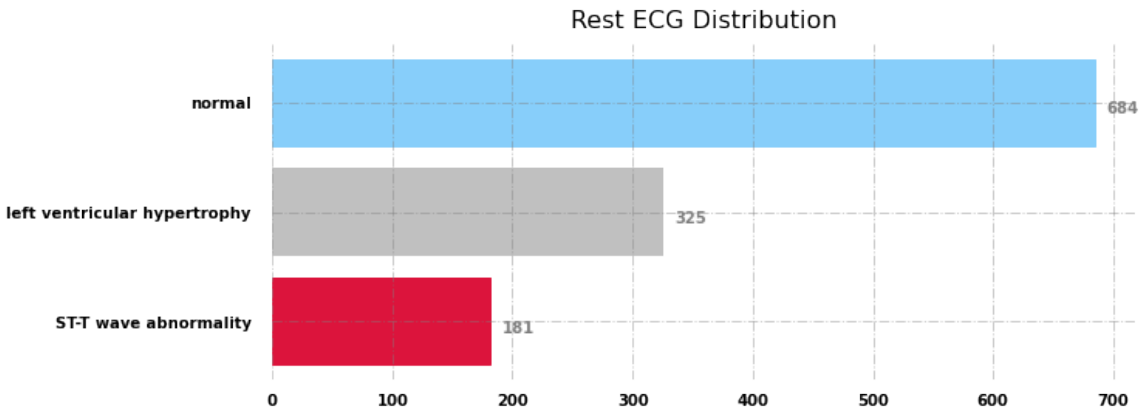


Figure 3.5: Rest ECG distribution

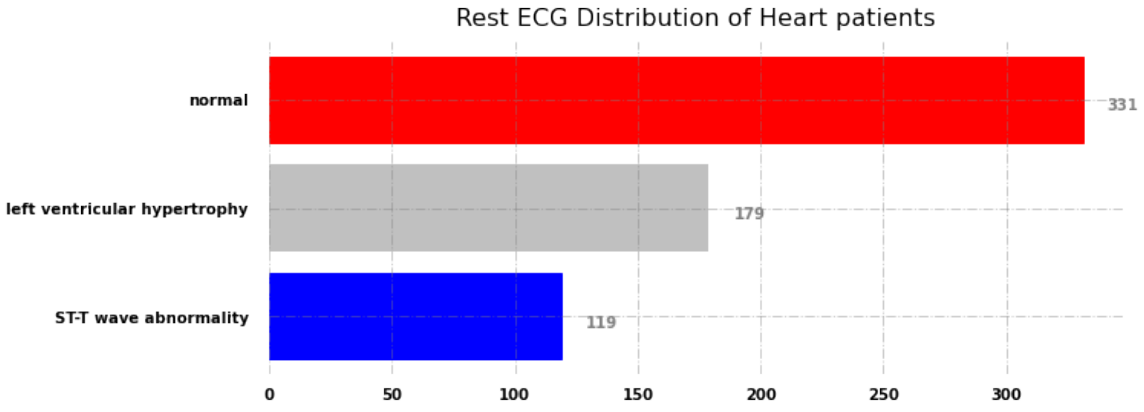


Figure 3.6: Rest ECG distribution of CAD patients

Target	0	1
rest_ecg		
ST-T wave abnormality	11.05	18.92
left ventricular hypertrophy	26.02	28.46
normal	62.92	52.62

Table 3.4: target VS rest ecg

3.3 Conclusion

The dataset that was used and a thorough example of exploratory data analysis (EDA) are presented in this chapter's introduction. We shall go into the technique section to talk about the method that was employed in a later chapter.

Chapter 4

Research Methodology

The primary component of thesis that will be covered in this chapter is research methods. This chapter also focuses on the data processing techniques that are used; later, several learning methods will be discussed.

4.1 Introduction

This section essentially describes some of the strategies that were used in this study to employ various machine learning algorithms. The methodology section, which describes the process to be followed, is the main component of research activity.

4.2 Data Preprocessing

Two methods of data processing are employed in this study. Data normalization and data transformation are two examples. This chapter's prior section covered data normalization and data transformation. We utilized a dummy function to modify the data, turning the categorical variables into index variables. Additionally, data normalization maintains the continuous value inside a [0-1] standard range.

4.3 Machine Learning Algorithms

Numerous applications and areas of research in business and academia are being transformed by machine learning. Machine learning affects every area of our lives, from programmatic

adverts that anticipate our wants before we ever think about them to voice assistants who help us organize appointments, verify the calendars, and listen to music. In the research, we adopted a supervised learning strategy.Two categories are used to categorize these learning difficulties. Regression is the first, and grouping is the second. We attempt to map input variables to output variables to around 25 continuous functions in a regression issue in order to predict outcomes within an infinite output. In a classification task, we are instead attempting to forecast the path that will lead to a discrete outcome. Additionally, we make an effort to categorize these input variables.

4.3.1 Regression

We must identify the subject's age based on the notion of a human picture.Rectilinear regression is one example of this method.

- estimating someone's age, race, or if a company's stock price will increase tomorrow
- determining whether a paper makes a UFO sighting claim?

4.3.2 Classification

In a patient with a tumor, we must determine whether the tumor is malignant or benign.Examples of algorithms include Support Vector Machine, Naive Bayes, Logistic Regression, Decision Trees, and others.

- identifying a person's gender based on the way they write
- home price forecasting sponsored region
- determining if the monsoons will be typical the following year
- estimating how many copies of a music record will be sold in the coming month

The machine learning algorithms that were applied in this research will now be discussed.

4.4 Support Vector Machine

We used the support vector machine approach in our research, which is typically used by the majority of academics. In this part, a brief overview of support vector machines will be covered.

4.4.1 Introduction of Support Vector Machine

A computer algorithm called an SVM (Support Vector Machine) learns to mark objects through examples[38].SVM is a prominent ML approach that separates data points using a hyperplane.Boser, Guyon, and Vapnik employed the Support Vector Machine for the first time in COLT-92 in 1992. SVMs are a group of supervised learning algorithms that are used for classification and regression[39].They are classified as generalized linear classifiers.Support Vector Machine uses a classification and regression prediction method to increase predictive accuracy.The hypothesis space of linear functions is frequently utilized in support vector machines in high-dimensional feature spaces. It originally piqued the interest of the NIPS group, who are now a crucial component of the ml research. In a handwriting identification task, SVM achieves accuracy comparable to complex neural networks with intricate features when pixel maps are utilized as data[40].With a focus on pattern recognition and regression, it is also employed in a range of applications, including face analysis and handwriting analysis.Vapnik[41] laid the groundwork for support vector machines, which have grown in prominence due to a number of promising characteristics like empirical effectiveness.The formula makes use of the Structural Risk Minimization (SRM) idea, which has been proved to perform better than the classic Empirical Risk Minimization (ERM) idea employed by conventional neural networks[42].While ERM minimizes the error on the training data, SRM minimizes the estimated risk. This discrepancy improves SVM's generalization capabilities,which is the main goal of statistical learning. SVM were initially developed to address classificaitons issues, but more recently they have been extended to address regression[43].

4.4.2 Support Vector Machine

Every machine learning expert should be familiar with SVM, which is a somewhat straightforward algorithm. Because the support vector machine reaches effective accuracy with less processing power, many researchers prefer it. It is frequently employed in classification and regression issues. It is, nevertheless, frequently applied to classification objectives. For categorizing n-dimensional space, the SVM algorithm generates the shortest decision boundary or line. The optimum judgment boundary is a hyperplane. Support Vector Machine selects the acute points/vectors that aid in creating the hyperplane. These severe cases are referred to as SVM. let's take a look at the diagram below, that defines how to define two distinct categories

using a decision boundary or hyperplane:

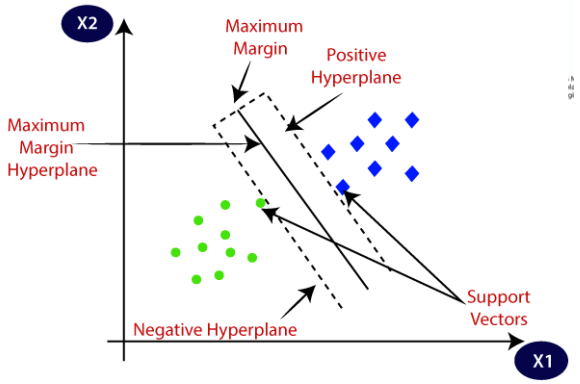


Figure 4.1: Support Vector Machine(SVM)[2]

4.4.3 Support Vector Machine:Example

We may understand Support Vector Machines by utilizing the KNN classifier as an example.If we see an unusual cat withsome dog-like qualities and want a model that can properly assess whether its a cat or a dog, we may use the SVM approach to develop such a model.Lets put our model to the test with this unique creature after first training it with a ton of images of cats and dogs so that it can learn about their varied traits.The acute example of cat and dog may be understood as a result of the support vector's development of a preference border between these two data (cat and dog) and preference for extreme scenarios (support vectors). On the basis of the support vectors definition, it will be classified as a cat. Consider the illustration below:

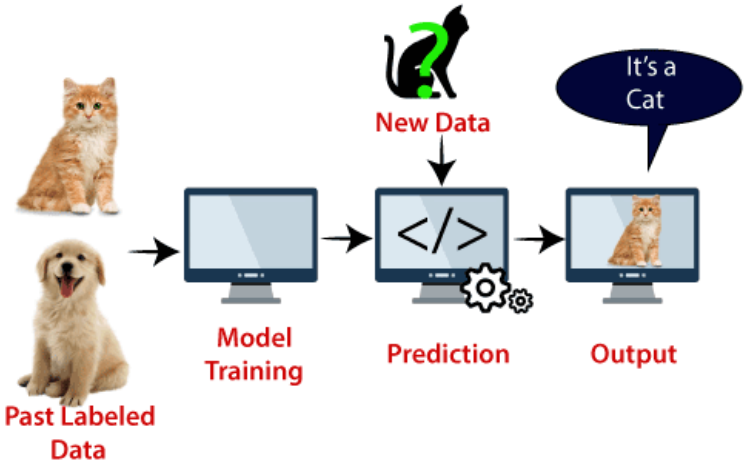


Figure 4.2: Support Vector Machine Example[2]

4.4.4 Support Vector Machine Varieties

Support vector machines are classified into two types.Linear Support Vector Machines and Non Linear Support Vector Machines are the two varieties.

4.4.4.1 Linear Support Vector Machine

Linearly separable data is defined as data that can be divided into two groups by a single line, and the classifier used for this sort of data is known as a linear support vector machine.

4.4.4.2 Non-linear Support Vector Machine

In other words, if a dataset could be classified using a line, it is classed as non-linear data, and the classifier used is Non-linear SVM.

4.4.5 Hyperplanes and Support Vector Machine

Hyperplane decision boundaries help in the categorization of data objects. The groups assigned to the data point either side of the of the hyperplane are also different.The quantity of functions also determines the hyperplane's size.The hyperplane is only a line if the input features are limited to just two. If there are three input features, a two-dimensional plane is the hyperplane. It gets impossible to picture once there are 3 features. The support vector's data points are

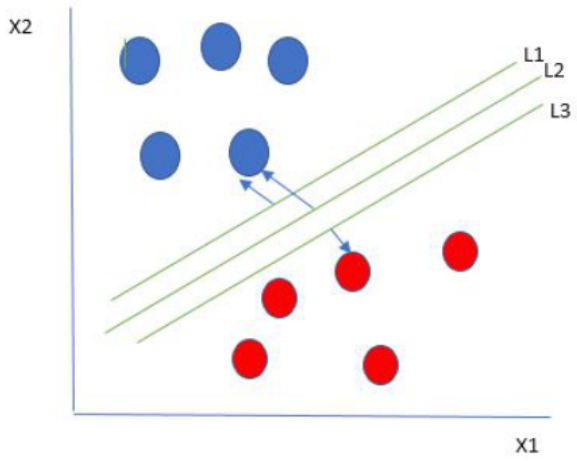


Figure 4.3: Hyperline that separate Data[3]

located closer to the hyperplane, which affects the orientation and placement of the hyperplane.

Using these support vectors boosts the classifier's margin. In the event that these support vectors are eliminated, the hyperplane's location will change.

4.4.6 How does the Support Vector Machine work?

Support Both linear and nonlinear samples can be processed by a vector machine. Below is an explanation of these two types of data for your consideration:

4.4.6.1 linear Support Vector Machine

It is customary to use an example to describe how the Support Vector Machine algorithm operates. Assume we have a dataset with two features called x1 and x2 and two tags (green and blue). A classifier is required to determine if the pair of coordinates (x1, x2) is green or blue. The example that follows is: We can easily identify these two groups with just a line as this is a

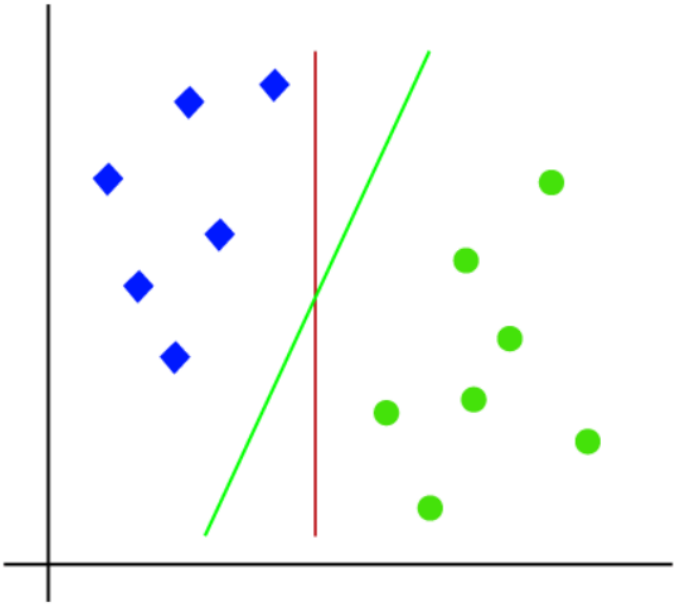


Figure 4.4: Hyperplane within 2D space[2]

two-dimensional space. These groupings are, however, frequently split by a number of lines.As a result, the SVM method helps in choosing the simplest boundary or region for the decision; this area or boundary is known as a hyperplane.This algorithm locates the intersection of the lines from both groups.These points are referred to as support vectors.

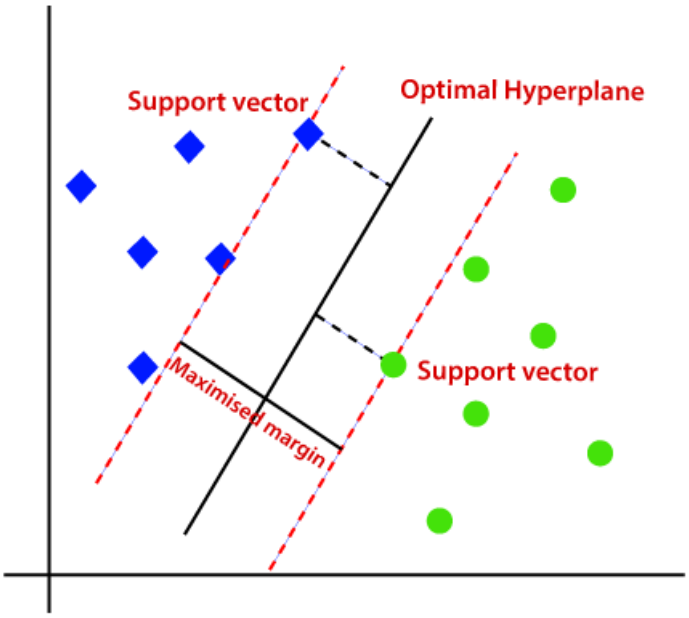


Figure 4.5: Optimal Hyperplane[2]

4.4.6.2 For non-linear Support Vector Machine

If the data are linearly ordered, we can use a line to divide them, with the exception of non-linear data, which cannot be divided by a single line.Look for the following example:

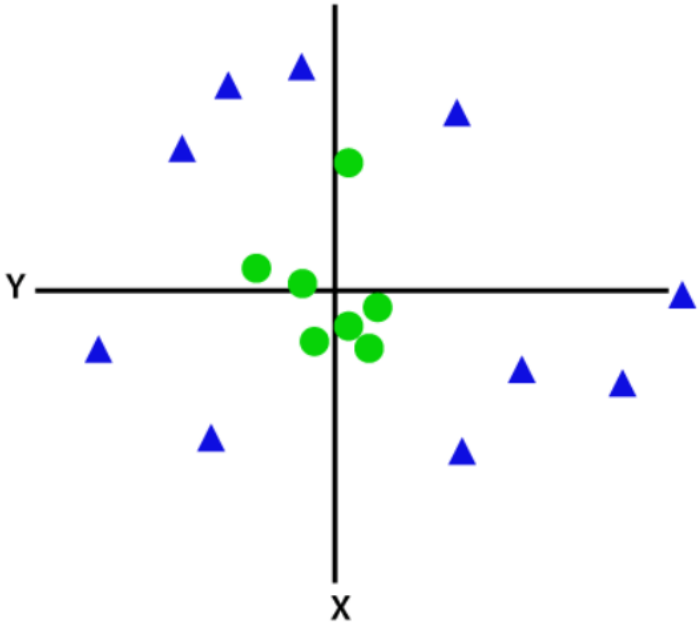


Figure 4.6: Non-Linear data in 2D space[2]

We'd want to add a new dimension to distinguish these data points.We've utilized two dimensions for linear data, x and y, therefore for non-linear data, we'll add a third dimension, z. A

popular formula for calculating it is z=x2 +y2.After adding the dimension, Support Vector Machine splits datasets into classes as shown below. Find the following illustration:

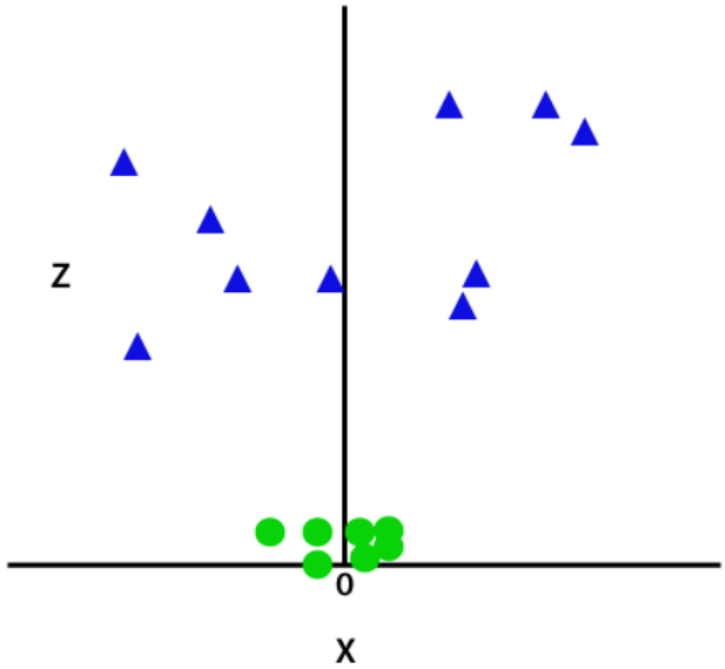


Figure 4.7: 3D sample[[2]

It appears to be a sort of plane parallel to the x-axis because we are in three dimensions. Using z=1 to convert it to 2D space now, the plane will appear as :

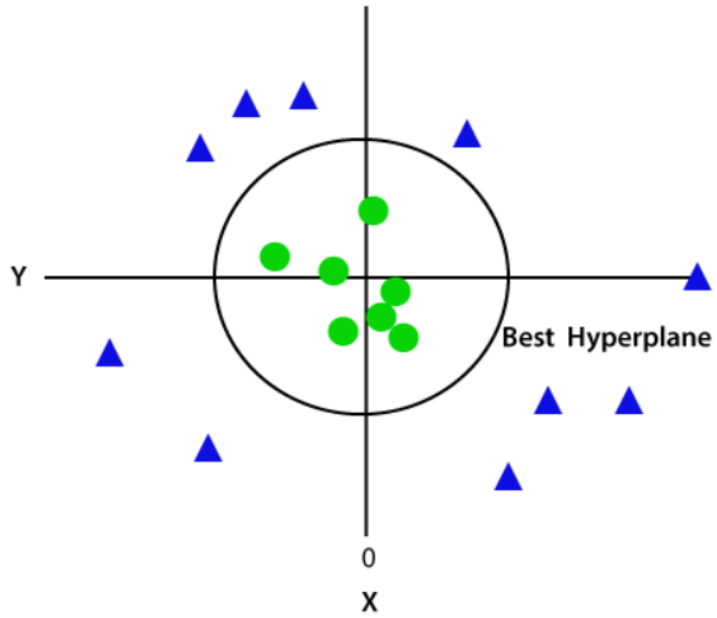


Figure 4.8: 3D figure shown in 2D space[[2]

Consequently, while dealing with non-linear data, we obtain a perimeter of radius 1.

4.4.7 Support Vector Machine Intuition with a High Margin

In logistic regression, the sigmoid function is used to constrict the value of the output of the linear function to the range [0, 1]. The value is labeled 1 if it exceeds the threshold value (0.5), else it is labeled 0. In SVM, we obtain the output of a linear function and, depending on whether it is greater than 1 or less than 1, we allocate it to one class or the other. Because the edge values change to 1 and -1, we obtain the reinforcement range ([-1,1]) that governs as margin.

4.4.8 SVM with Radial Basis Function(RBF) kernel

You're working on a Machine Learning technique like Support Vector Machines for non-linear datasets and you can't seem to figure out the proper feature transform or the right kernel to utilize.Well, fear not because Radial Basis Function (RBF) Kernel is your rescuer.Due to its resemblance to the Gaussian distribution, RBF kernels are one of the most extensively used kernels and the most general kind of kernelization. The RBF kernel function computes the closeness or degree of similarity of two points X1 and X2.This kernel may be mathematically described as follows:

$$K(X1, X2) = \exp -\frac{\|X1 - X2\|^2}{2\sigma^2} \tag{4.1}$$

where,

- 'σ' is the variance and our hyperparameter
- ||X1 - X2|| is the Euclidean (L2-norm) Distance between two points X1 and X2

4.5 Random Forest

This section briefly describes about random forest. Random forest is usually employed by many researchers.

4.5.1 Introduction of Random Forest

random Forest is a typical machine learning approach that belongs to the supervised learning methodology.It may be used for both classification and regression issues in machine learning.It

is based on ensemble learning, which is the act of combining several classifiers to solve a complicated issue and enhance model performance."Random Forest is a classifier that consists of a number of decision trees on different subsets of the supplied dataset and takes the average to increase the predicted accuracy of that dataset," as the name indicates.Instead than depending on a single decision tree, the random forest takes into account each tree's forecast and predicts the final output based on the majority vote of predictions. The increasing number of trees in the forest increases precision and eliminates the issue of overfitting[4].

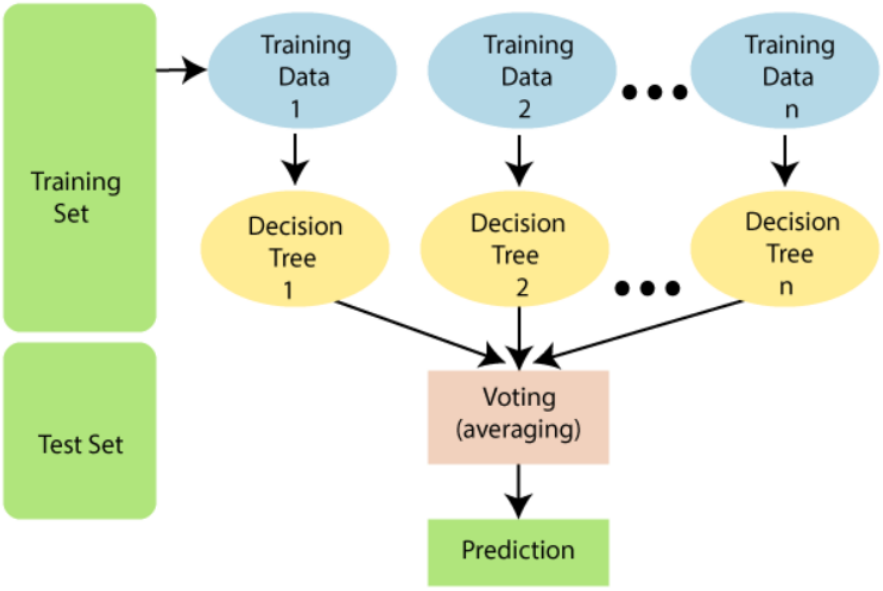


Figure 4.9: Random forest algorithm[4]

4.5.2 Random Forest Assumptions

Because the random forest mixes many trees to forecast the class of the dataset, some decision trees may predict the correct output while others may not.However, the trees as a whole properly predict the outcome.As a result, two assumptions for a more accurate Random forest classifier are as follows:

- Actual values in the dataset's feature variable should be used here so that the classification may make precise predictions rather than assumptions
- The forecasts of each tree must have very low correlations[4]

4.5.3 Working of Random Forest Algorithm

To understand how the random forest works, we must first analyze the ensemble approach.The term ”ensemble” simply refers to the merging of several models.As a consequence, rather than a single model, a set of models is used to make predictions. Ensemble employs two categories of techniques:

- **Bagging:** The sample training data is replaced with a new training subset, and the outcome is determined by majority vote. Consider Random Forest.
- **Boosting:** It converts weak learners into strong ones by developing successive models, with the final model having the highest accuracy. ADA BOOST and XG BOOST, for instance.

As previously stated, Random Forest operates on the Bagging concept.Bagging, also known as Bootstrap Aggregation, is the ensemble approach used by random forest. Bagging is used to choose a random sample from the data collection. As a consequence, each model is built using the samples (Bootstrap Samples) provided by the Original Data, using a technique called as row sampling.This step of row sampling with replacement is referred known as the bootstrap. Currently, each model is being trained individually and delivering results.After combining the results of all the models, a majority vote is used to make the final conclusion. Aggregation is the process of combining all of the results and obtaining a final outcome based on a majority vote.

4.5.4 Grid Search for Hyperparameter Tuning a Random Forest Classifier

A machine learning algorithm’s learning and behavior are controlled by hyperparameters. Hyperparameters are model characteristics (such as the number of estimators for an ensemble model) that we must pre-set, in contrast to internal parameters (such as coefficients, etc.), which the algorithm automatically optimizes during model training.It is difficult to determine the ideal hyperparameter configuration. The appropriate hyperparameters are frequently impossible to predict in advance. Therefore, it is necessary to do numerous tests with various parameters in order to obtain a great model.This can take a long time to perform manually.This article introduces the grid search strategy, a rigorous method for pinpointing a machine learning model’s

ideal hyperparameters.Using a grid of preset hyperparameters, grid search examines all conceivable permutations and delivers the model variant that gives the best results (the search space). We'll examine how this operates in the case of hyperparameterizing a classification model in the sections that follow. To achieve this, we will create and improve a random decision forest in Python that categorizes passengers on the Titanic as survivors or not[5].

The grid search method is based on a very straightforward concept. The objective is to experiment with different configurations of our model until we are happy with the outcome. Grid search is thorough in that it examines every possible combination of a parameter grid. The parameter grid and the given parameters determine the number of model variations. We must supply the following data to the grid search algorithm:

- The hyperparameters we wish to set (e.g., tree depth)
- For each hyperparameter a range of values (e.g., [50, 100, 150])
- a performance metric, so the algorithm can choose how to assess performance (e.g., accuracy for a classification model[5])

The illustration below shows a sample parameter grid:

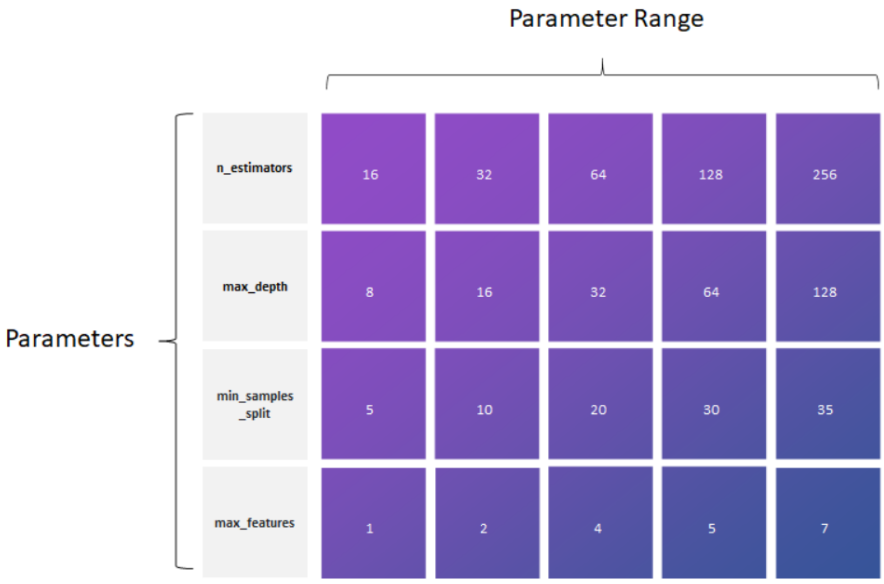


Figure 4.10: A sample parameter grid with four hyperparameters for adjusting a random decision forest[5]

4.6 Naive Bayes Classifier

The introduction of the Naive Bayes Classifier is followed by a conclusion in this section.A brief explanation of Naive Bayes will be discussed in between them. Naive Bayes

4.6.1 Introduction of Naive Bayes Classifier

Inferential statistics and a variety of sophisticated machine learning models depend on the Bayes theorem. As a logical method for revising ideas' likelihood in light of fresh evidence, Bayesian reasoning has a significant place in science[44].In order to answer problems for which frequentist statistical methods have not been defined, we can apply Bayesian analysis. In actuality, the frequentist paradigm prohibits ascribing a likelihood to a theory. The goal of this text is to give a concise and mathematically precise exposition of the origins of Bayesian statistics:The Bayes theorem[45], a straightforward yet effective machine learning technique, is the foundation of the Naive Bayes classifier.Despite its simplicity, Naive Bayes may also perform better than more sophisticated categorization techniques[46].In 2004, the Bayesian classification dilemma showed that there are some explanations for the egregiously unreasonable efficacy of the Naive Bayes classifiers[47].However, a detailed comparison of different classification techniques in 2006 showed that more modern strategies, like boosted trees or random forests[48], perform better than Bayes classification.

4.6.2 What is the significance of the name Naive Bayes?

The two words Naive and Bayes, which make up the Naive Bayes classifier algorithm, are explained as follows:

Naive: Because it thinks the existence of one function has no bearing on the appearance of other features, it is known as naive.A red, spherical, sweet fruit is referred to as an apple if the fruit's shape and flavor are used to categorize it.As a result, each feature functions independently of the others to classify an apple.

Bayes: It is based on the Bayes theorem, hence the name Bayes.

4.6.3 Bayes Theorem

The Bayes' theorem, often known as the Bayes' Rule or Bayes' rule, is a mathematical technique for evaluating the probability of a hypothesis based on past information.It is determined by the likelihood. The following is the formula for the Bayes theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B} \tag{4.2}$$

where,

- **P(A|B) is Posterior probability:**Probability of hypothesis A on the observed event B.
- **P(B|A) is Likelihood probability:** Probability of the evidence as long as the probability of a hypothesis is true.
- **P(A) is Prior Probability:** Probability of hypothesis before monitoring the evidence.
- **P(B) is Marginal Probability:** Probability of Evidence

4.6.4 Sorts of Naive Bayes Model

The following list of three different types of Naive Bayes models:

4.6.4.1 Gaussian

The Gaussian model assumes a standard distribution for the features.This suggests that the model assumes that if predictors take continuous values rather than discrete ones, their values are samples from the normal distribution.

4.6.4.2 Multinomial

The Multinomial Nave Bayes classifier is employed when the data is dispersed over multiple variables. It is mostly used to resolve problems with document classification, such as determining which category a specific document falls under, such as Sports, Politics, or Education. Based on the frequency of phrases, the classifier's predictors make predictions.

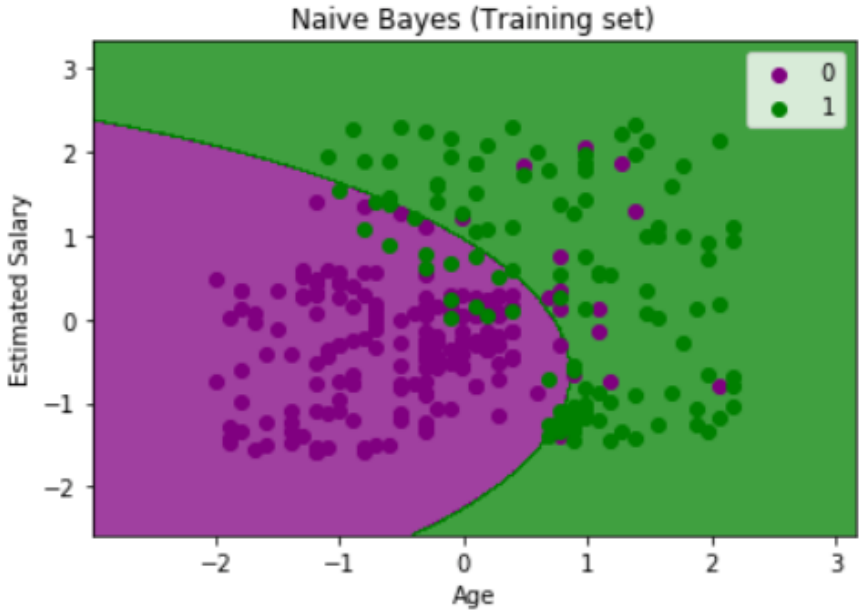


Figure 4.11: Naive Bayes Classifier as Gaussian Curve[6]

4.6.4.3 Bernoulli

The predictor variables in the Bernoulli classifier differ from those in the Multinomial classifier in that they are distinct Boolean variables. such example if a specific word appears in a text or not. This model is renowned for its ability to classify documents.

4.7 Neural Network

This section begins with introduction of Neural Network and ends with a conclusion. In between of them brief description of Neural Network are going to be described.

4.7.1 Introduction of Neural Network

A neural network, also known as an artificial neural network learning algorithm or simply a neural net, is a computer learning system that employs a network of functions to interpret and convert a desired output, often in another form, from a single kind of data input. The notion of the artificial neural network was founded on human biology and how neurons interact together in the human brain to absorb information from the senses.A range of tools and approaches, including neural networks, are used by machine learning algorithms.

The neural network itself may be used as a component in many different machine learning approaches to translate complicated data inputs into a language that computers can understand.

Today, neural networks are utilized to handle a wide range of real-world problems, including voice and image recognition, spam email filtering, finance, and medical diagnosis, to name a few. [7].

4.7.2 How Does a Neural Network Work?

Typically, neural network-based machine learning algorithms do not need to be developed with specific rules that indicate what to expect from the input.Instead, the neural network learning algorithm learns by examining a large number of labeled samples (i.e., data with ”answers”) and using this answer key to identify what attributes of the input are necessary to produce the desired output.After a sufficient number of instances have been processed, the neural network may begin processing new, untested inputs and dependably deliver dependable outputs.

Because the computer learns by experience, the more instances and various inputs it encounters, the more accurate the results are typically. An example will help you grasp this concept better. Consider the ”simple” problem of determining whether or not an image includes a cat. While it is very simple for a human to understand, it is far more challenging to teach a computer to recognize a cat in an image using conventional techniques. Writing programming that takes into consideration every conceivable variation in how a cat can seem in a photograph is practically difficult. However, the computer can utilize a generalized technique to analyze the content in an image by employing machine learning, and more especially neural networks. The neural network can begin to recognize trends that exist throughout the many, many examples that it analyses and classify images by their similarity by using several layers of functions to breakdown the image into data points and information that a computer can use.

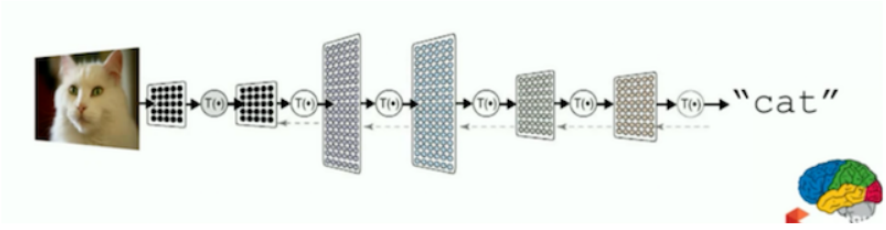


Figure 4.12: How Neural Network looks[7]

After analyzing numerous training samples of cat photos, the algorithm has developed a model

of the factors that should be taken into account when determining whether or not a cat is in the image. The neural network compares the data points regarding a new image to its model, which is based on all past assessments, while evaluating the new image. Then, based on how well the image matches the model, it determines whether or not the image contains a cat using some straightforward statistics[7]. The neural network in this illustration is made up of the layers of functions that are present between the input and the output. The neural network is a little bit more intricate in reality than it appears in the image above. Although the interaction between layers is significantly better depicted in the graphic below, there are numerous variants in the connections between nodes, or artificial neurons:

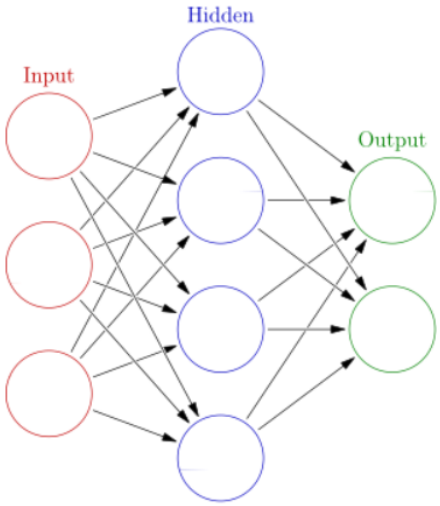


Figure 4.13: basic structure of Neural Network[7]

4.7.3 Multilayer Perceptron

The multi layer perceptron augments the feed forward neural network (MLP).It is made up of three distinct types of layers: input layer, output layer, and hidden layer.The input layer is where the processing signal is received.The output layer completes tasks such as categorization and prediction.The MLP's true computational engine is made up of an arbitrary number of hidden layers situated between the input and output layers.Data travels in the forward direction from the input to the output layer of an MLP, similar to a feed forward network.The MLP's neurons are trained using the back propagation learning process.Because MLPs are designed to approximate any continuous function, they may tackle problems that are not linearly separable.MLP's key applications include pattern categorization, recognition, prediction, and approximation[49].

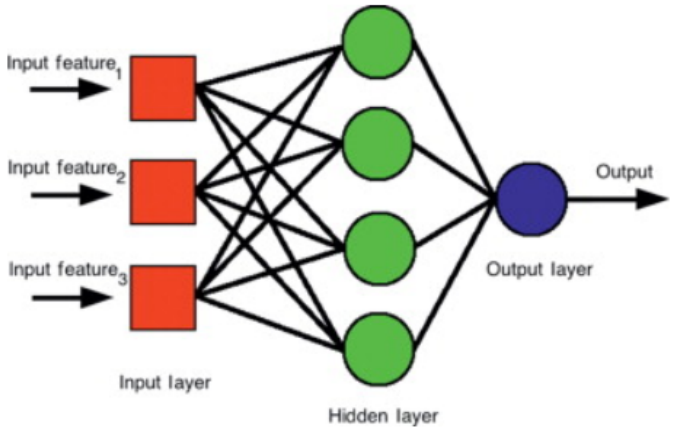


Figure 4.14: Multilayer perceptron[7]

4.7.4 Activation Function

It is merely a thing function that you use to obtain a node's output.It also goes by the name Transfer Function[50].

4.7.4.1 Why we use Activation functions with Neural Networks?

It is used to identify the neural network's output, which is either yes or no.The values acquired are mapped between 0 and 1 or -1 and 1, and so on (depending upon the function) [50]. The two main categories of activation functions are:

- Linear Activation Function
- Non-linear Activation Functions

4.7.4.2 ReLU (Rectified Linear Unit) Activation Function

Currently, the ReLU is the most widely used activation function on a worldwide scale.

Because it is used in almost all convolutional neural networks and deep learning systems.

He ReLU has been somewhat adjusted (from bottom).

f(z) = zero when z is less than zero, and f(z) equals z when z is greater than or equal to zero. **Range:**[From 0 to infinity] The function and its derivative are both monotonic. However, any negative values are quickly converted to zero, reducing the model's ability to properly fit or train from the data.This implies that any negative input to the ReLU activation function becomes zero in the graph instantly, which has an effect on the final graph by inappropriately mapping the negative values.

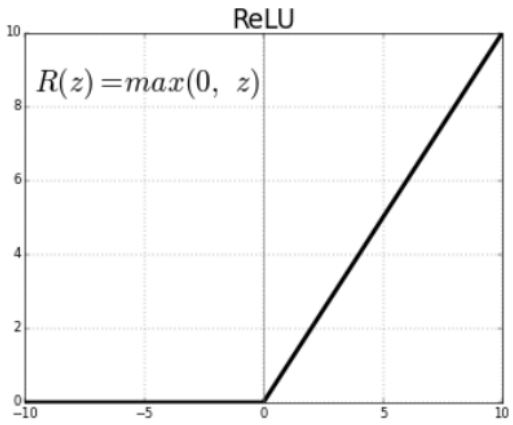


Figure 4.15: ReLU function[7]

4.8 Conclusion

The SVM method is the most basic data modeling methodology.They integrate generalization and dimensionality control.The kernel mapping offers a common base for several of the most often used model designs, allowing for comparisons[51]. The random forest is a categorization method composed of many decision trees.By employing bagging and feature randomization while producing each individual tree, it seeks to construct an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.Naive Bayes one of the simplest and quickest machine learning algorithms for predicting a collection of datasets, Naive Bayes is commonly employed for both binary and multi-class classifications.It beats other algorithms in single-class predictions as well as in multi-class predictions.As a consequence, it is typically included as an option in classification algorithms. Neural networks are effective at predicting time series because they can learn from examples alone, eliminating the requirement for additional information that could generate more confusion than benefit.The ability to generalize and noise resistance of neural networks. here, we've attempted to describe the strategies we used in our research projects.In this area, processing and machine learning algorithms are covered. In comparison to the thesis work, this chapter is important.

Chapter 5

Implementation

5.1 Introduction

The topic of methodology was covered in the section prior, and in this chapter we will demonstrate the use of some of the techniques and algorithms covered in that area. First, the implementation flow chart is provided below.

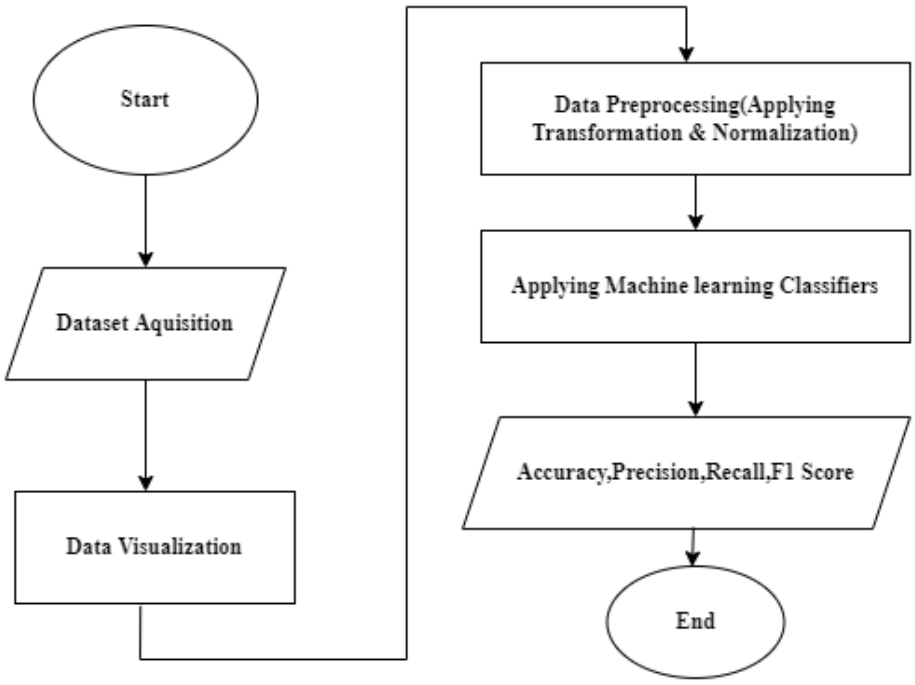


Figure 5.1: Workflow for Implementation

5.2 Data Acquisition

Kaggle is used to collect the dataset.The chapter 2 contains a description of the dataset.The dataset is ready for viewing after being acquired.

5.3 Data Visualization

Better visualization models can be created using machine learning algorithms, which are built to understand prior data and apply their conclusions to fresh information. This makes it possible to make datasets more descriptive, improving the notion for displaying knowledge. The dataset indicates that 629 cases have CAD and 561 instances are normal, or 53% of the instances have CAD disease and 47% of the instances are normal.

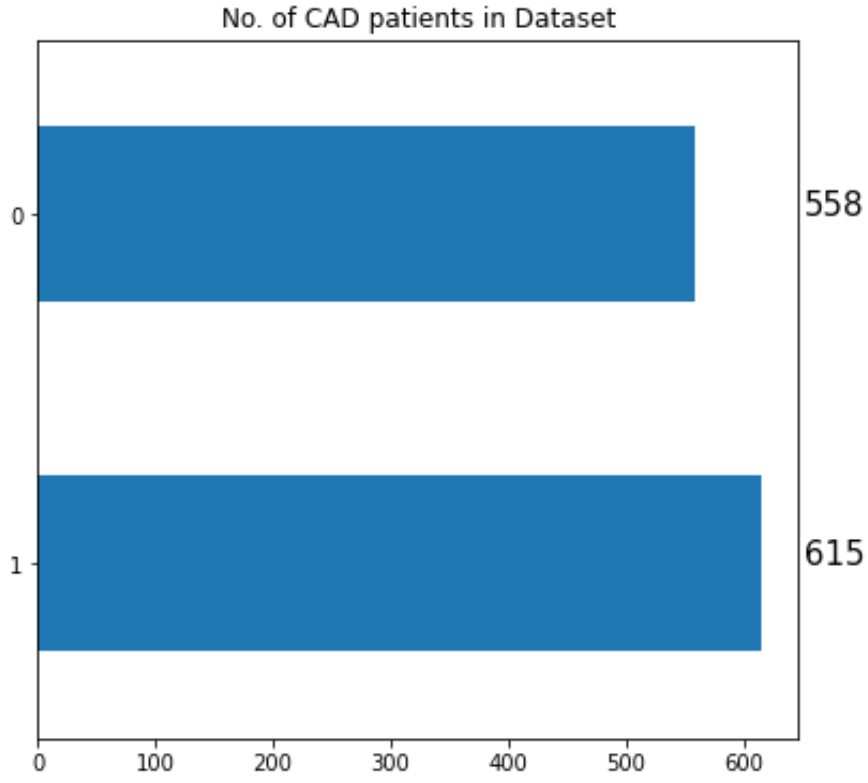


Figure 5.2: Number of CAD patients in dataset

In the above figure, 0 means Normal patients and 1 means CAD patients.

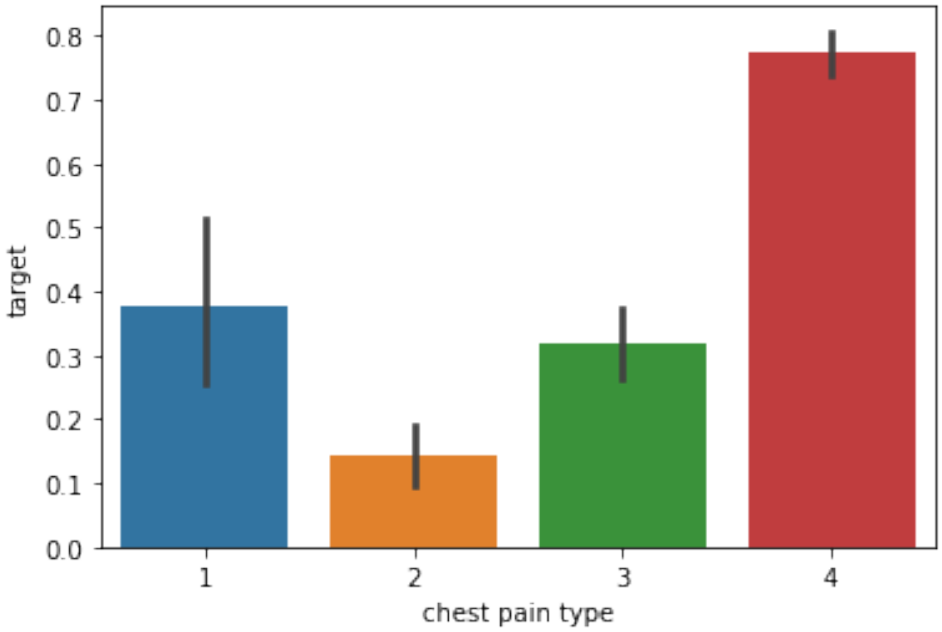


Figure 5.3: Barplot of Chest Pain type

The barplot of various types of chest pain is displayed in the above figure.

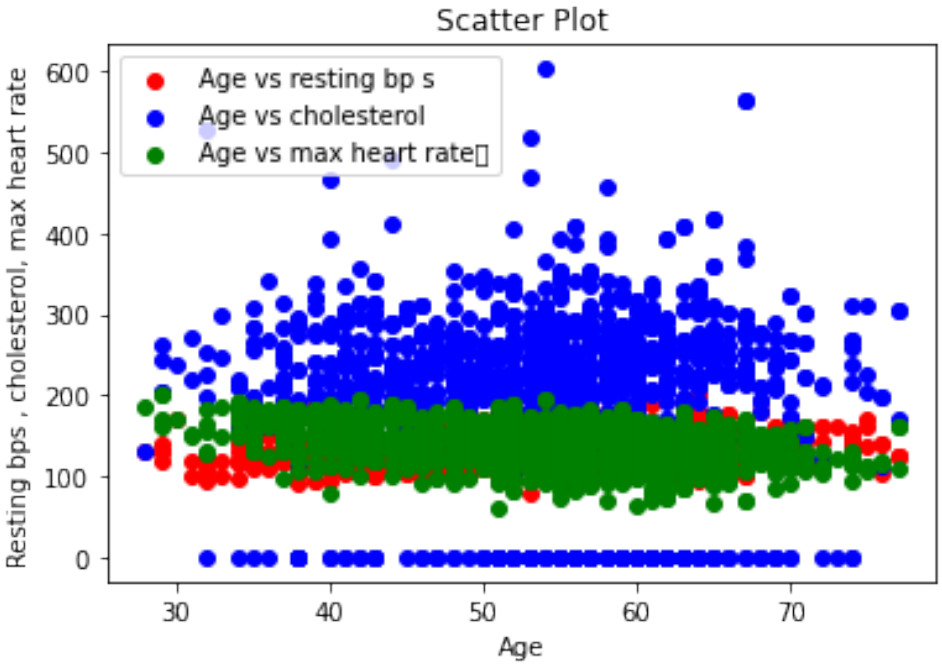


Figure 5.4: Scatter Plot

This scatterplot shows age vs resting bps, cholesterol and max heart rate. X

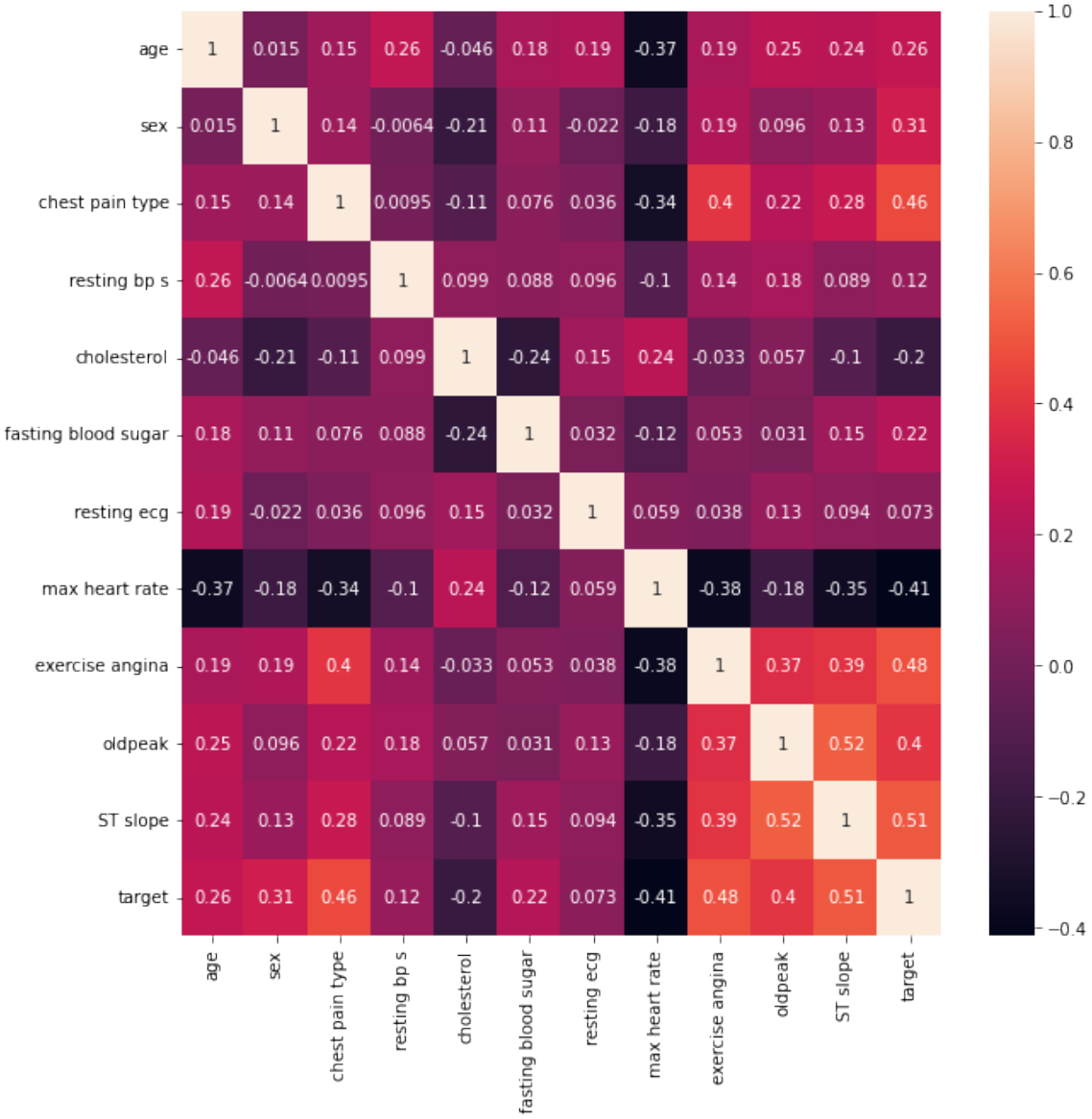


Figure 5.5: Correlation Matrix

The above figure shows the correlation among the attributes.

5.4 Data Preprocessing

In the data preprocessing phase, data transformation and normalization are conducted. Prior to using machine learning algorithms, it is crucial. Predictions become more accurate as a result.

5.5 Applying Machine Learning Classifiers

The primary component of every thesis study is the application of classifier algorithms. The accuracy of prediction cannot be obtained without classifier implementation.

5.5.1 Implementation Step

- Import the dataset
- Investigate the details to determine how they appear.
- Pre-Process the data
- Split the dataset into labels and characteristics.
- Train the Support Vector Machine /Random Forest/Naive Bayes/Neural Networks
- Make predictions
- Analyze the algorithm's output. algorithm

5.6 Confusion Matrix , Accuracy, Precision, Recall and F1 score

We are aware that a classification problem's prediction summary is a confusion matrix. As a result, each class weakens and reduces the number of correct and incorrect predictions. Confusion matrix demonstrates the methods by which a classification model can become perplexed when making predictions. It provides insight into both the errors and, more importantly, the kind of errors that are being made. Four components make up the confusion matrix table: True Positive, True Negative, False Positive, and False Negative.

- Any phenomenon will be True Positive if a model predicts a real value and that particular value is also true (TP).
- Any phenomenon will be True Negative if any model successfully predicts a negative value (TN).
- Actual value in false positive phenomena is ”No,” while predicted value is ”Yes” (FP).

- In false negative phenomena, the true answer is ”Yes,” but the answer that is predicted is ”No” (FN).

These four parameters will be used to help determine Accuracy, Precision, Recall, and F1 score.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{5.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{5.3}$$

$$F1score = 2 * \frac{(Recall * Precision)}{(Recall + Precision)} \tag{5.4}$$

5.7 Conclusion

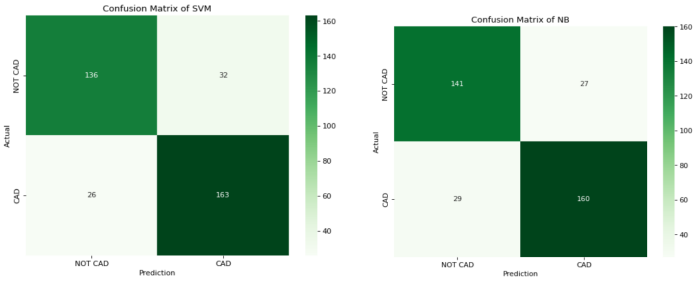
This chapter can be a key component of the study. One cannot accomplish their objective without good implementation. The researcher will not be successful in his endeavors unless the following approaches are correctly applied. We appropriately applied machine learning algorithms in order to obtain better results.

Chapter 6

Result and Performence Analysis

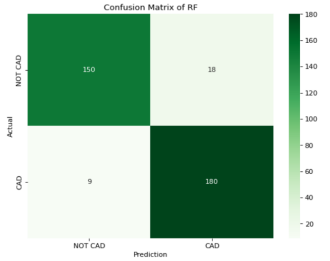
6.1 Result

For any researcher, this is the portion they want most. It will be an accomplishment for a researcher if the outcome is significantly better than a base publication or earlier research. The confusion matrix, which was covered before in this book, is presented first. The final display includes accuracy, precision, recall, and f1 score. As a result, the findings of this study are as follows:

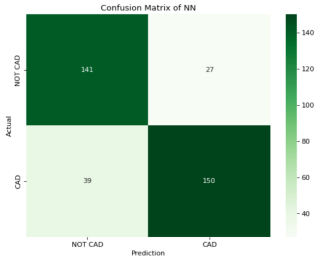


(a) SVM Confusion Matrix

(b) NB Confusion Matrix



(c) RF Confusion Matrix



(d) NN Confusion Matrix

Figure 6.1: Confusion Matrix

6.2 Performance Analysis

Table 6.1: Precision,Recall F1 score of Machine Learning Classifiers

Classifiers	Precision	Recall	F1 Score
Support Vector Machine	0.97	0.69	0.81
Naive Bayes	0.84	0.85	0.85
Random Forest	0.97	0.91	0.94
Neural Network	0.82	0.83	0.82

Table 6.2: Accuracy Table

Classifiers	Previous Accuracy	Present Accuracy
Support Vector Machine	0.86	0.85
Naive Bayes	0.82	0.84
Random Forest	0.90	0.935
Neural Network	-	0.81

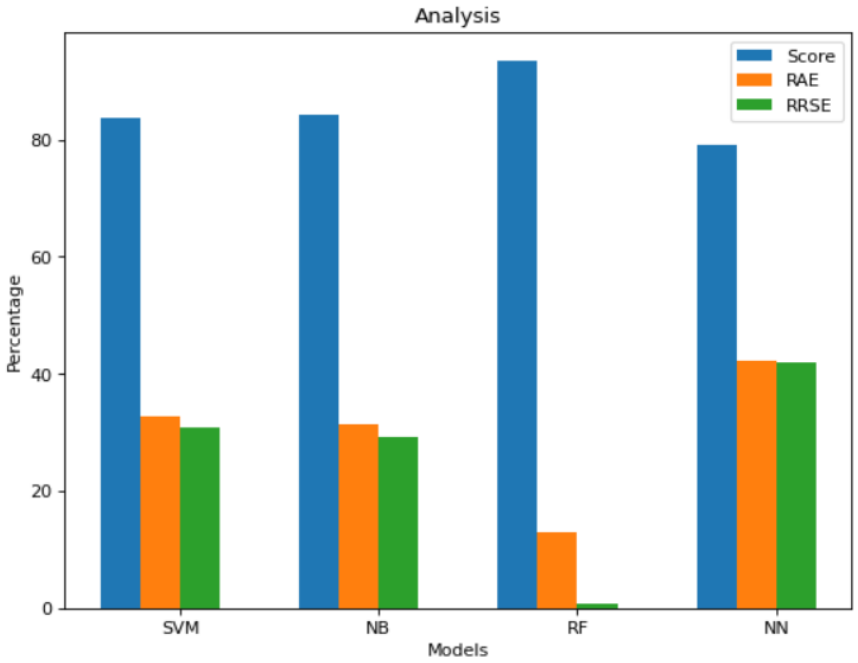


Figure 6.2: Analysis

So, finally, we have the outcomes we were hoping for.This data clearly shows that Random Forest provides greater accuracy than the other three classifiers.Furthermore, Naive Bayes provides more accurate results than Support Vector Machine.

Chapter 7

Conclusion and Future Works

7.1 Introduction

analysis of the experiments were discussed in the preceding chapter. This chapter will begin with a discussion of a synopsis of the entire thesis research, followed by a discussion of the thesis' limitations.The following section provides suggestions for prospective future works, and the chapter concludes with a final conclusion.

7.2 Thesis Summary

The main purposes of this thesis are to identify and analyze CAD disease in an early stage and to select the best suitable classifier to predict CAD.This thesis will result in a productive resolution in life science, particularly in the field of heart disease.Experts have now penetrated every aspect of our lives and developed our everyday routines.We attempted to see patients' datasets and forecast the results based on this thesis.

7.3 Limitations

our work has some limits.These algorithms' accuracy ought to be 100 percent of the time.As a result, the predictions made by these classifier systems have failed to predict accurately.

7.4 Future Works

This study is frequently expanded for use in subsequent research that will employ different machine learning algorithms and processing approaches.Additionally, this might help both patients and doctors comprehend the nature of the CAD condition. Combining different classifiers might produce better results. A new dataset with more features on Neural network could demonstrate improved accuracy.

7.5 Conclusion

Four alternative classification algorithms are briefly described in this thesis. The outcomes of those algorithms differ after processing compared to before. Algorithms don't provide the best accuracy when the dataset isn't trained adequately. As a result, these models will provide the highest level of accuracy when the data has been appropriately trained.

REFERENCES

[1] CDC, "Coronary artery disease, available at: https://www.cdc.gov/heartdisease/coronary_ad.htm."

[2] Javapoint, "Support vector machine algorithm,available at:https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm."

[3] Geeeksforgeeks, "Support vector machine,available at:https://www.geeksforgeeks.org/support-vector-machine-algorithm/."

[4] Tutorialpoints, "Random forest algorithm,available at: https://www.javatpoint.com/machine-learning-random-forest-algorithm/."

[5] F. Follonier, "Hyperparameter tuning a random forest classifier using grid search in python,available at: https://www.relataly.com/hyperparameter-tuning-with-grid-search/2261/."

[6] JavaTpoint, "Naïve bayes classifier algorithm, available: https://www.javatpoint.com/machine-learning-naive-bayes-classifier."

[7] DeepAi, "Neural network,what is a neural network?, available: https://deepai.org/machine-learning-glossary-and-terms/neural-network."

[8] "Healthy environment, healthy heart (internet). jakarta: Ministry of health, republic of indonesia; 2014. available at:http://www.depkes.go.id/article/ view/201410080002/lingkungan-sehat-jantung-sehat.html.."

[9] WHO, "Cardiovascular diseases (cvds),available at: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)."

[10] T. F. Members, G. Montalescot, U. Sechtem, S. Achenbach, F. Andreotti, C. Arden, A. Budaj, R. Bugiardini, F. Crea, T. Cuisset, *et al.*, "2013 esc guidelines on the management of stable coronary artery disease: the task force on the management of stable coronary artery

disease of the european society of cardiology," *European heart journal*, vol. 34, no. 38, pp. 2949–3003, 2013.

[11] K. B. Fuster V, "Institute of medicine (us) committee on preventing the global epidemic of cardiovascular disease: Meeting the challenges in developing countries; fuster v, kelly bb, editors. promoting cardiovascular health in the developing world: A critical challenge to achieve global health. washington (dc): National academies press (us); 2010. available from: https://www.ncbi.nlm.nih.gov/books/nbk45693/ doi: 10.17226/12815."

[12] T. Y. Wah, R. Gopal Raj, U. Iqbal, *et al.*, "Automated diagnosis of coronary artery disease: a review and workflow," *Cardiology research and practice*, vol. 2018, 2018.

[13] N. D. Wong, "Epidemiological studies of chd and the evolution of preventive cardiology," *Nature Reviews Cardiology*, vol. 11, no. 5, pp. 276–289, 2014.

[14] Z. Shehzad, U. R. Hameed, Z. Hira, G. Sadia, K. Nayab, S. Saira, N. Sadia, S. Mutahira, and A. Bibi, "Heart attack pervasiveness along with associated risk factors in district headquarter hospital in karak, khyber pakhtunkhwa, pakistan," *Journal of Coastal Life Medicine*, vol. 4, no. 11, pp. 896–897, 2016.

[15] S. Mendis, P. Puska, B. Norrving, W. H. Organization, *et al.*, *Global atlas on cardiovascular disease prevention and control*. World Health Organization, 2011.

[16] P. K. Mehta, J. Wei, and N. K. Wenger, "Ischemic heart disease in women: a focus on risk factors," *Trends in cardiovascular medicine*, vol. 25, no. 2, pp. 140–151, 2015.

[17] F. J. Charlson, A. E. Moran, G. Freedman, R. E. Norman, N. J. Stapelberg, A. J. Baxter, T. Vos, and H. A. Whiteford, "The contribution of major depression to the global burden of ischemic heart disease: a comparative risk assessment," *BMC medicine*, vol. 11, no. 1, pp. 1–12, 2013.

[18] WHO, "World health rankings , availble at : https://www.worldlifeexpectancy.com/bangladesh-coronary-heart-disease."

[19] F. Farzadfar, "Cardiovascular disease risk prediction models: challenges and perspectives," *The LANCET Global Health*, vol. 7, no. 10, pp. e1288–e1289, 2019.

[20] “W. 2019, heart disease cases soaring in bangladesh, the daily star, 2020. available: https://www.thedailystar.net/world-heart-day-2019/heart-disease-cases-soaring-in-ban gladesh-1806820..”

[21] medicalnewstoday, “What to know about coronary artery disease.available at:https://www.medicalnewstoday.com/articles/184130.”

[22] “What is coronary artery disease.available at: https://www.webmd.com/heart-disease/coronary-artery-disease.”

[23] D. Bzdok, M. Krzywinski, and N. Altman, “Machine learning: supervised methods,” *Nature methods*, vol. 15, no. 1, p. 5, 2018.

[24] “What is machine learning? a definition - expert system, expert.ai, 2020.available :https://www.expert.ai/blog/machine-learning-definition/.”

[25] R. Sathya, A. Abraham, *et al.*, “Comparison of supervised and unsupervised learning algorithms for pattern classification,” *International Journal of Advanced Research in Artificial Intelligence*, vol. 2, no. 2, pp. 34–38, 2013.

[26] E. Diederichs, “Reinforcement learning–a technical introduction,” *Journal of Autonomous Intelligence*, vol. 2, no. 2, p. 25, 2019.

[27] S. Kalyanakrishnan and P. Stone, “Characterizing reinforcement learning methods through parameterized learning problems,” *Machine Learning*, vol. 84, no. 1, pp. 205–247, 2011.

[28] X. Zhou and M. Belkin, “Semi-supervised learning,” in *Academic Press Library in Signal Processing*, vol. 1, pp. 1239–1269, Elsevier, 2014.

[29] F. D. Purba, J. A. Hunfeld, T. S. Fitriana, A. Iskandarsyah, S. S. Sadarjoen, J. J. Busschbach, and J. Passchier, “Living in uncertainty due to floods and pollution: the health status and quality of life of people living on an unhealthy riverbank,” *BMC public health*, vol. 18, no. 1, pp. 1–11, 2018.

[30] B. Jin, C. Che, Z. Liu, S. Zhang, X. Yin, and X. Wei, “Predicting the risk of heart failure with ehr sequential data modeling,” *Ieee Access*, vol. 6, pp. 9256–9261, 2018.

[31] H. D. Masethe and M. A. Masethe, "Prediction of heart disease using classification algorithms," in *Proceedings of the world Congress on Engineering and computer Science*, vol. 2, pp. 25–29, 2014.

[32] A. K. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," *Neural Computing and Applications*, vol. 29, no. 10, pp. 685–693, 2018.

[33] C.-s. M. Wu, M. Badshah, and V. Bhagwat, "Heart disease prediction using data mining techniques," in *Proceedings of the 2019 2nd international conference on data science and information technology*, pp. 7–11, 2019.

[34] M. Siddhartha, "Heart disease dataset(comprehensive). available at :https://www.kaggle.com/datasets/sid321axn/heart-statlog-cleveland-hungary-final."

[35] Purba, "Stacked ensemble for heart disease classification. available at :https://www.kaggle.com/code/purba01/stacked-ensemble-for-heart-disease-classification."

[36] WIKI, "Exploratory data analysis. available at : https://en.wikipedia.org/wiki/exploratory_data_analysis."

[37] D. N. Bhamri, "Rise in cases of asymptomatic heart attacks amongst middle aged people . available at : https://www.maxhealthcare.in/blogs/rise-cases-asymptomatic-heart-attacks-amongst-middle-aged-people."

[38] B. E. Boser, "isabelle m," *Guyon, and Vladimir N. Vapnik."A Training Algorithm for Optimal Margin Classifiers." COLT*, vol. 92, 1992.

[39] WIKI, "support vector machine. available at : https://en.wikipedia.org/wiki/support_vector_machine.."

[40] WIKI, "A. moore, awm tutorial page, cs.cmu.edu, 2020. available at : http://www.cs.cmu.edu/ awm/tutorials.html.."

[41] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 1999.

[42] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[43] V. Vapnik, S. Golowich, and A. Smola, "Support vector method for function approximation, regression estimation and signal processing," *Advances in neural information processing systems*, vol. 9, 1996.

[44] D. A. Berry, *Statistics: a Bayesian perspective*. No. 04; QA279. 5, B4., 1996.

[45] Tutorialpoints, "Naive bayes classifier in machine learning,available at:https://www.javatpoint.com/machine-learning-naive-bayes-classifier."

[46] I. Karimov, S. Jafarova, M. Zeynalli, S. Rustamov, A. Z. Adamov, and A. Babakhanov, "Transformation, analysis and visualization of distributed temperature sensing data generated by oil wells," in *2020 IEEE 14th International Conference on Application of Information and Communication Technologies (AICT)*, pp. 1–7, IEEE, 2020.

[47] H. H. Zhang, "Cs.unaive bayes.ca, 2020. [online]. available: http://www.cs.unaive bayes.ca/profs/hzhang/."

[48] cornelledu, "Naivee bayes,available:https://www.cs.cornell.edu/čaruana/ctp/ct.papers /caruana.icml06.pdf.."

[49] P. Raj and P. E. David, *The Digital Twin Paradigm for Smarter Systems and Environments: The Industry Use Cases*. Academic Press, 2020.

[50] S. SHARMA, "Activation functions in neural networks, available: https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6."

[51] N. Cristianini and T. De Bie, *Support vector machines*. Hodder Arnold, 2005.