Cardiovascular Disease Prediction Model Using Machine Learning Algorithm

by

MD Shamsul Arefin Mirdha
ID: 12101037
Anannya Ahmed
ID: 14301067

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
September 2022

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

MD Shamsul Arefin Mirdha
ID: 12101037

Anannya Ahmed
ID: 14301067

Approval

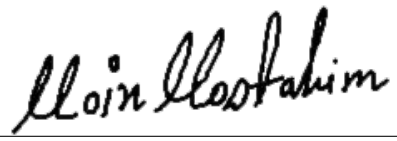
The thesis/project titled "Cardiovascular Disease Prediction Model Using Machine Learning Algorithm" submitted by

1. MD Shamsul Arefin Mirdha (12101037)
2. Anannya Ahmed (14301067)

Of Summer, 2022 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on September 25, 2022.

Examining Committee:

Supervisor:
(Member)



Moin Mostakim
Senior Lecturer
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Md. Golam Rabiul Alam, PhD
Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi
Chairperson
Department of Computer Science and Engineering
Brac University

Abstract

This research uses machine learning to anticipate and detect the symptoms of specific diseases after examining some of the important elements of these diseases in order to better understand them and develop new and better treatment techniques. This study uses machine learning and generated data sets to evaluate and categorize the signs and symptoms of heart diseases. We'd like to see whether we can improve individual disease prediction processes so that we can predict cardiovascular diseases and their modalities more accurately. Therefore, the aim of this study is also to develop a more diversified model from the existing ones. We are focusing on cardiovascular diseases, which is among the world's top causes of death. Multiple machine learning (ML) algorithms are being used more frequently to predict cardiovascular disease. We want to evaluate and describe how well ML algorithms generally forecast cardiovascular illnesses. This research analyzes the classification of cardiovascular disease using machine learning methods including Random Forest (RF), Logistic Regression, Decision Tree, Naïve Bayes, Linear Algorithm, Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Neural Network. We anticipate finding effective and efficient results that will aid in better diagnosing these cardiovascular diseases and also will help us for developing better treatment procedures.

Keywords: Cardiovascular Disease; Machine Learning; Random Forest (RF); Logistic Regression; Decision Tree; Naïve Bayes; Linear Algorithm; Support Vector Machine (SVM); K-Nearest Neighbor.

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.

Secondly, to our Advisor Mr. Moin Mostakim Sir for his kind support and advice in our work. He helped us whenever we needed help. His valuable advice, critical criticism and active supervision encouraged me to sharpen my research methodology and was instrumental in shaping my professional outlook

And finally to our parents without their throughout support it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

AI Artificial Intelligence

$COPD$ Chronic Obstructive Pulmonary Disease

CVD Cardiovascular Disease

KNN K-Nearest Neighbor

ML Machine Learning

SVM Support Vector Machine

WHO World Health Organization

Chapter 1

Introduction

Healthcare is an unavoidable aspect of human life. The impact of diseases on human health is increasing significantly as a result of many changes in our environment, such as climate change, changes in people's lives, and other causes. As a result, the ratio of sick individuals has gone through the roof. People are dying from chronic obstructive pulmonary disease (COPD) at a substantially higher rate than from any other disease. Cardiovascular diseases is one of the deadliest amongst the other chronic obstructive pulmonary (COPD).

The body's circulatory system, which also includes the lungs, is mostly made up of the muscular organ known as the heart, which pumps blood into the body. The cardiovascular system is also composed of a network of blood vessels, such as capillaries, arteries, and veins. All over the body, blood is transported via these blood vessels. Different types of heart disorders, often known as cardiovascular diseases, are brought along by irregularities from the heart's regular blood flow (CVD). Plaque buildup in the arteries causes an obstruction in the blood flow to the heart, which results in a heart attack. A thrombus in an artery that prevents blood from reaching the brain can result in a stroke.

Unhealthy eating, inactivity, smoking, and alcohol use are the greatest behavioral risk factors for cardiovascular disease and stroke. Based on their medical features, such as gender, age, chest discomfort, fasting blood sugar level, etc, this study seeks to determine whether a patient is likely to be diagnosed with any cardiovascular cardiac abnormalities.

1.1 Thoughts behind the Prediction Model

World's one of the leading cause of death Cardiovascular Disease, has long been a major public health concern, inflicting enormous socioeconomic damage on patients, families, and countries every year [11] . According to current studies, chronic diseases cause approximately 3 million deaths each year. With the help of modern technologies, this alarming rate can be decreased. Reducing these high mortality rate can possibly happen if the detection of the disease can be done accurately.

Every disease, as we all know, has its own set of therapy options. So, if we can accurately diagnose each patient's disease, the treatment process will be considerably more accurate as well. If the detection process is not done accurately, then there are high possibilities that one type of disease will be mistaken with another disease. It is challenging for practitioners to diagnose because the symptoms are similar to those associated with other disorders and could be mistaken with aging-related signs. So, by using Machine Learning, we can develop a model that can be used to detect diseases accurately such as cardiovascular disease.

As we all know, we cannot deny our healthcare system's reliance on technology. Many technological innovations had already took place over the last few decades, which is advancing our healthcare systems in a variety of ways. There are different types of technological mechanism used in medical applications. Among all other techniques, machine learning (ML) is one of the best technique that plays an essential role in anticipating diseases accurately and providing doctors and caregivers with the ability to deliver better treatment strategies accordingly.

Making computer models that can access and use data on their own to learn is the core objective of machine learning. The need for trustworthy medical treatments for people with various chronic conditions is one of the most important motivations for using machine learning (ML) into pharmaceuticals. Modern healthcare systems are sophisticated and smart because they are driven by data and models. Which is needed not only for detecting diseases like cardiovascular diseases but also for professional who are practicing medicines.

1.2 Aims and Objectives

In a study it shows that approximate 17.5 million people passed away in 2012 from coronary disease, which means that it accounts for 31% of all mortality globally. Furthermore, the number of people who die from cardiovascular disease is increasing year after year. By 2030, it is anticipated to grow to a population of more than 23.6 million. According to research published in January 2017, cardiovascular infections are the leading cause of death worldwide. As a result, researchers all over the world are working to introduce new technologies and appropriate technological mechanisms to medical science in order to tackle these diseases. Some examples of CVD includes heart attack, stroke, heart failure, arrhythmia, and heart valve disorders.

Modern technological methods help medical healthcare systems by forecasting a patient's current condition utilizing a variety of ML methods, including Naive Bayes, Decision Trees, SVM, Random Forests, Logistic Regression, and KNN. These algorithms can be applied for predicting the disease, after which we can perform analysis and decide which ones are ideal for the project we are doing. These methods enable us to produce the appropriate health alerts by using a patient's relevant features. In order to predict and determine the accuracy of the given data set, machine learning incorporates a variety of classifiers from supervised, unsupervised, and ensemble learning. These techniques will be helping us for collecting and gathering information and then will predict any patient's real time conditions and the disease they are suffering from.

The proposed method for predicting heart disease will improve medical care while costing less. This study provides us with important information that can assist us anticipate who will get heart disease. The dataset was obtained from the Kaggle repository, and Python was used to create the model.

Chapter 2

Problem Statement

Cardiovascular disease can also be managed by some lifestyle changes, by adding some medication plan also in some cases by doing a surgery like any other diseases. Any symptoms related to cardiovascular disease can be reduced and the functioning of heart can also be improved by taking a proper care and medication.

For our Thesis research, we are going to use Machine Learning (ML) to create a prediction model for cardiovascular diseases. In our paper we are analyzing different symptoms of cardiovascular disease from different data sets using algorithms of Machine Learning (ML) and generate outputs. So by analyzing the outputs that are generated, it will be easier for us to understand the pattern of this disease and how the patients are affected because of different factors associated with this disease.

Chapter 3

Research Objectives

Cardiovascular diseases affect millions of individuals around the globe. Sometimes they suffer without having a clear understanding of these diseases. The goal of this thesis paper is to identify cardiovascular diseases using data gathered from online platforms such as kaggle. Using these data sets, we will attempt to discover the pattern of these diseases.

This study will go through the causes of these diseases as well as try to predict them using alternate methodologies. Our goal is to create a model that uses Machine Learning Algorithms to predict and identify the kind cardiovascular diseases based on some of its major characteristics.

The main motive of our research is to create a Machine Learning-based prototype Health Care Prediction System. From a data set of cardiac disorders, the System can find and extract hidden knowledge related to diseases. It also aids in lowering treatment costs by offering efficient treatments.

This research presents the results in tabular and PDF versions to improve visualization and enhance interpretation. It can respond to comprehensive queries for disease detection, which helps medical professionals make wise clinical decisions that conventional decision support systems cannot.

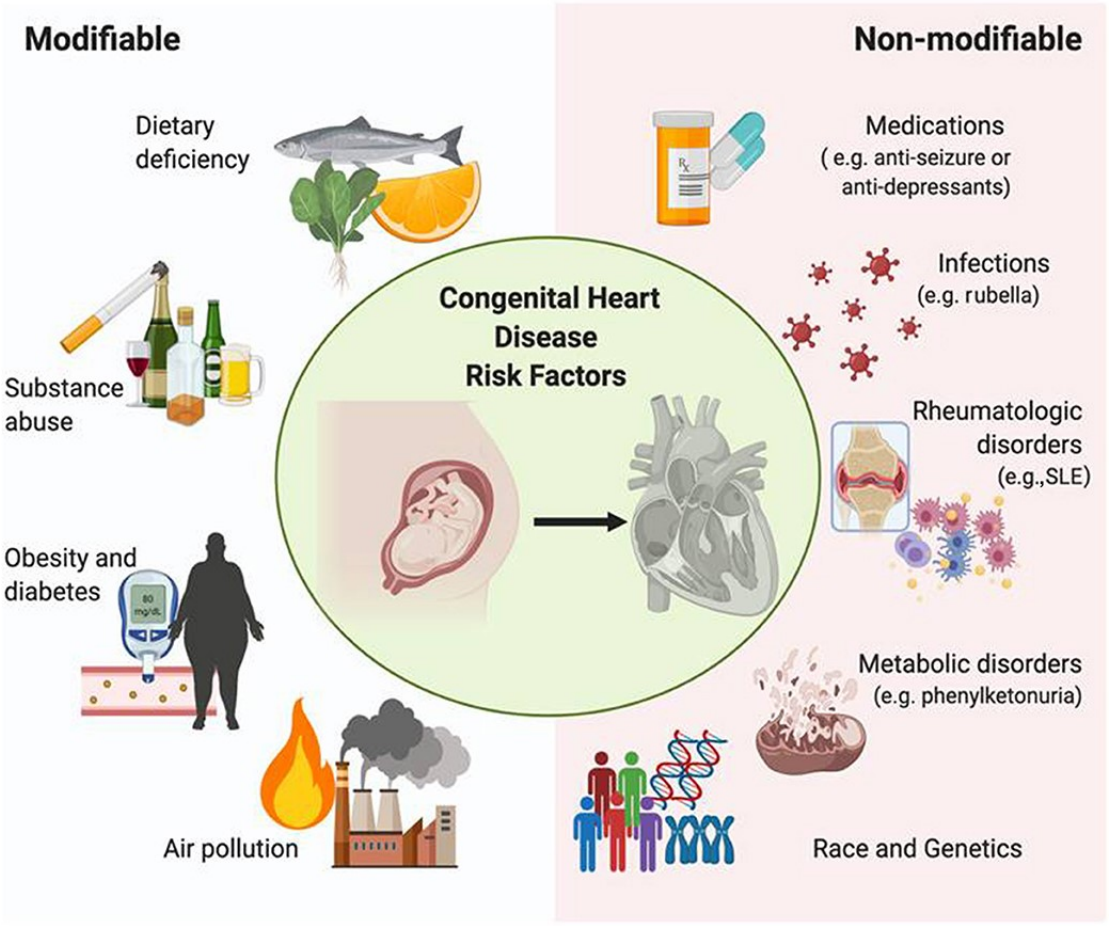


Figure 3.1: Modifiable and Non-modifiable Risk Factors of Cardiovascular Disease

Chapter 4

Risk Factors of Cardiovascular Disease

Cardiovascular disease kills more than 17 million people worldwide each year, according to the World Health Organization. The most common form of cardiovascular illness, coronary artery disease, is the leading cause of death in the world today. Major and contributory risk factors are separated into two groups. Heart disease risk is increased by a number of major risk factors. Heart disease can be exacerbated by risk factors that contribute to it. The higher your risk of developing heart disease, the more risk factors you have. Some risk factors can be altered, addressed, or changed, while others cannot. However, you can minimize your risk of heart disease by managing as many risk factors as possible through lifestyle changes, medications, or both.

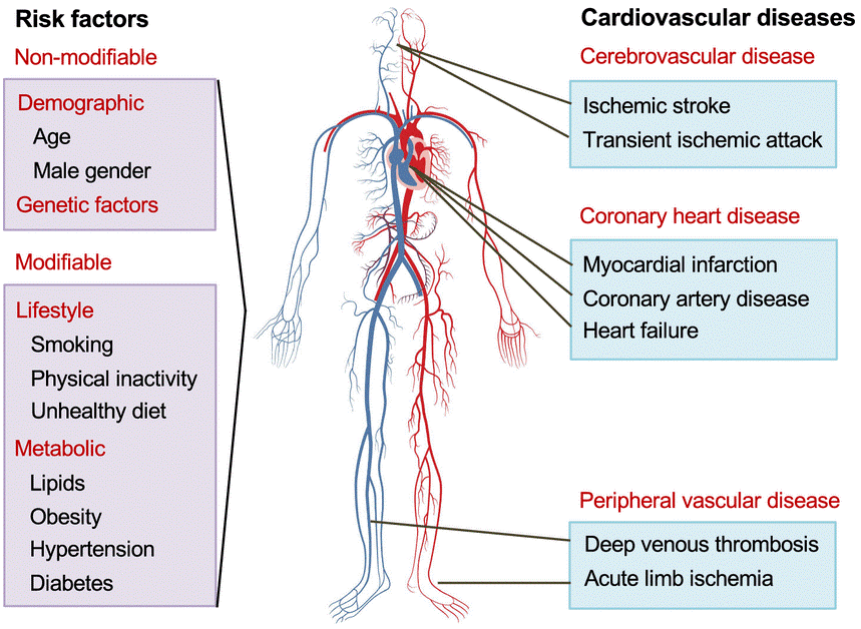


Figure 4.1: Classification of Cardiovascular Disease Risk Factors

4.1 Major Risk Factors

High blood pressure raises the risk of heart disease, heart attack and stroke. Obesity, smoking or having high blood cholesterol levels in addition to high blood pressure increases your risk of heart disease or stroke significantly. Blood pressure varies with exercise and age, but a healthy adult's resting blood pressure should be 120/80.

High cholesterol level is a major risk factor for cardiovascular disease. One of the primary causes of heart disease is high blood cholesterol. Cholesterol is a fat-like substance carried in your blood and found in all of your body's cells. The cholesterol your body requires for cell membrane creation and hormone manufacturing is produced entirely by your liver. Extra cholesterol enters your body when you consume animal-based foods or foods high in saturated fat.

Diabetes is another risk factor. Heart disease is the major cause of mortality among diabetics, particularly those with adult-onset or Type 2 diabetes (also known as non-insulin-dependent diabetes). Diabetes is more common in some racial and ethnic groups (African Americans, Hispanics, Asian and Pacific Islanders, and Native Americans). According to the American Heart Association, 65 percent of diabetic individuals die from cardiovascular disease. If anyone has diabetes, they should already be under the supervision of a doctor because effective blood sugar control can lower your risk. If you suspect you have diabetes but are unsure, consult your doctor for tests.

Overweight and Obesity Increased high cholesterol levels, high blood pressure, and diabetes, all are important risk factors for heart disease and can result from excess weight. A BMI of above 25 is considered to be overweight according to the National Heart, Lung, and Blood Institute (NHLBI). Obesity is defined as a number more than 30.

Another major risk factor for cardiovascular disease is smoking. Smoking cigarettes and tobacco raises the risk of lung cancer, but few people realize it also raises your risk of heart disease and peripheral arterial disease (disease in the vessels that supply blood to the arms and legs). Many of these deaths are caused by the effects of smoking on the heart and blood vessels. Smoking increases heart rate, narrows key arteries, and causes irregular pulses, all of which make your heart work harder. Smoking, which raises blood pressure, is another major risk factor. While nicotine is the most active element in cigarette smoke, other chemicals and compounds including tar and carbon monoxide can have a variety of effects on your heart.

Another factor is Heredity. Heart disease is a genetically transmitted disease. You are more likely to get heart disease if your parents or siblings had a heart or circulation problem before the age of 55 than someone who does not have a family history of heart disease. Risk factors such as high blood pressure, diabetes, and obesity can be passed down from generation to generation.

The influence of age is significant. Growing older is associated to heart disease. Heart disease kills around four out of every five persons over the age of 65. As we age, our hearts become less efficient. The heart's walls thicken, and arteries stiffen and harden, limiting the heart's ability to pump blood to the body's muscles. As a result of these changes, the risk of cardiovascular disease increases with age. Women are typically protected from heart disease until they reach menopause, when their sex hormones raise their risk.

4.2 Contributing Risk Factors

Stress is thought to play a role in the development of heart disease. Emotional stress, behavior habits, and socioeconomic level all have an impact on the risk of heart disease and heart attack.

Pills that prevent pregnancy is a contributing factor for cardiovascular disease. High-estrogen and progestin birth control pills have been related to an increased risk of heart disease and stroke, especially in women over 35 who smoke. Modern birth control tablets, on the other hand, have significantly lower quantities of hormones and are deemed safe for women under the age of 35 who do not smoke or have high blood pressure.

Alcohol is a significant contributing factor for cardiovascular disease. According to studies, people who drink moderate amounts of alcohol had a lower risk of heart disease than nondrinkers. Moderate consumption is defined as one to two drinks per day for men and one drink per day for women, according to specialists. Excessive alcohol use, on the other hand, can cause heart problems such as high blood pressure, stroke, irregular heartbeats, and cardiomyopathy (disease of the heart muscle). A normal drink contains 100-200 calories. Alcohol calories contribute to body fat gain, raising the risk of heart disease.

Chapter 5

Developing a Prediction Model

To learn from data in datasets, machine learning employs algorithms. They look for patterns, gain insight, make judgments, and assess their choices. Datasets are divided into two groups in machine learning. The training data is a subset of our actual dataset that is fed into the machine learning model to learn and uncover patterns. Our model is trained in this way.

After we have developed our machine learning model (using our training data), we will need unseen data to test it. This data is referred to as testing data, and it may be used to assess the performance and development of our algorithms' training, as well as change or optimize them for better outcomes.

5.1 Data Source

With consideration for their history of cardiac issues and in accordance with other medical conditions, an organized dataset of people had been chosen. Different diseases that affect the heart are referred to as heart diseases. The World Health Organization (WHO) reports that cardiovascular diseases are the leading cause of death among middle-aged persons. We use a data set that contains the medical histories of 1026 individual patients, all of varying ages. The medical features of the patient, such as age, resting blood pressure, fasting sugar level, etc., are included in this dataset, providing us with the much-needed information that enables us to determine whether or not a patient has been diagnosed with a heart condition.

This data set includes 14 medical characteristics for 1026 individuals that we can use to determine whether a patient is at risk for developing a cardiac condition or not. It also allows us to categorize patients into those who are at risk and those who are not at danger. This Heart Disease data set is taken from an online platform named Kaggle. Practicing data scientists and machine learning professionals can be found online at Kaggle, a division of Google LLC. Users can discover and share data sets on Kaggle, study and develop models in a web-based data science environment, collaborate with other data scientists and machine learning experts, and participate in competitions to address data science challenges.

5.2 Testing and Training Data Sets

Data are necessary for machine learning models. Even the most efficient algorithms may be rendered useless in the absence of a foundation of high-quality training data. Each and every ML algorithm requires data for input and output [8]. The data used to train an algorithm or machine learning model to anticipate the outcome that your model was designed to predict is known as training data. Your data will be modified with data labeling or annotation if you are employing supervised learning or some hybrid that incorporates that method. As a result, no component of machine learning is more important than high-quality training data.

The first data that is used to generate a machine learning model, from which the model develops and fine-tunes its rules, is referred to as "training data." The quality of this data has significant effects on the model's future development and establishes a strong standard for any applications that use the same training data in the future [5]. For the purpose of teaching a machine learning model, training data is a very huge data set. The selection of features that are important to certain business objectives is taught to prediction models that employ machine learning algorithms using training data. The labeled training data is used in supervised machine learning models. Unsupervised ML models are trained using unlabeled data [29].

The concept of using training data in machine learning systems is simple, but it is also fundamental to how these technologies operate. The machine learning model is created, to put it simply, using training data. It demonstrates what the desired result should look like. In order to fully comprehend the data set's features and improve performance, the model regularly analyzes the data [13]. The training set of data is the first collection of information used to teach a program how to employ technologies like Random Forest, KNN, SVM, and others to learn and produce complex outcomes. Additional data sets referred to as validation and testing sets may be used to complete it.

Figure 5.1: Training and Testing Data Set

On the other hand, the test data set is a subset of the training data set which is employed to provide an objective assessment of a final model [3]. Test data are used to evaluate the effectiveness of the algorithm you are using to train the algorithm, such as its accuracy or efficiency. You can use test data to determine how effectively your model, which is built on training data, can predict new outcomes.

An additional (or tertiary) data set used to test a machine learning algorithm after it has been trained on an initial training data set is known as a test set in machine learning. The concept is that, as opposed to being examined from a programming approach, predictive models always contain some form of unverified capacity that has to be tested [28]. If a model fits the test data set as well as it fits the training data set, then there has been a minimum level of over fitting. The traditional indicator of over-fitting is when the training data set fits the model more closely than the test data set [31].

The test results ought to;
1. Represent the entire original data set or a portion of it.
2. It ought to be big enough to make accurate predictions.

For the purpose of enhancing and validating machine learning models, training and test data are both significant.

In this below diagram it is shown how a testing and training data sets works on our prediction model;

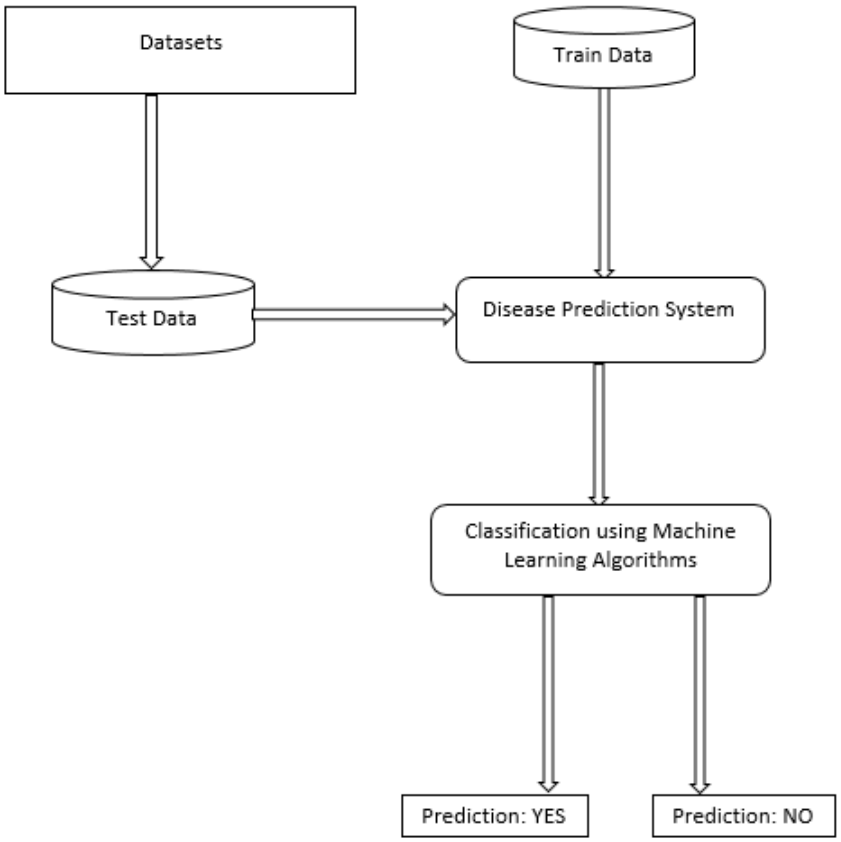


Figure 5.2: Testing and Training Data from Data Sets

The model would encounter each type of situation for a particular problem in the actual world, and the test data contains data for each of these scenarios. An ML project's test data set typically makes up 20–25% of the full original data set [22]. At this point, We can compare and contrast the training accuracy to the testing accuracy as well, or, more specifically, the accuracy of our model when applied to the test data set in comparison to the training data set. The model is considered to have over fitted if its accuracy on training data is higher than its accuracy on testing data.

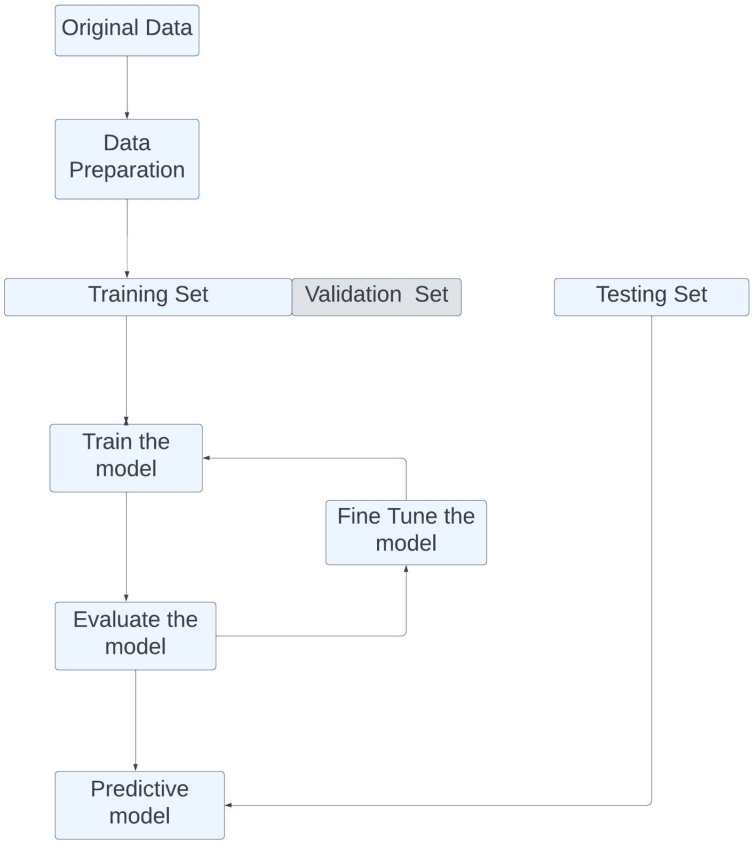


Figure 5.3: How Testing and Training Data Works

5.3 Designing a Prediction Model

The largest cause of death in the world is cardiovascular disease, which is also a serious public health issue. As a result, many of the treatment guidelines currently in use depend heavily on risk assessment. Risk management activities are also utilized to forecast the magnitude of future cardiovascular disease mortality and morbidity at the population level and among certain subgroups in order to alert policymakers and health authorities about these risks. Furthermore, risk prediction motivates people to improve their behaviors and lifestyles and to take their medications exactly as prescribed [6].

This article explores various machine learning algorithms. The algorithms employed in this paper are; Random Forest, Logistic Regression, Decision Tree, Naive Bayes, Support Vector Machine (SVM), K-Nearest Neighbor (KNN). This research includes a review of the cardiovascular disease data set, published articles and also scholarly journals. This research follows a procedure that incorporates actions that transform a provided data set into recognized data for the consumers' understanding. The steps in the proposed technique (as shown in Figure 5.4 Cardiovascular Disease Prediction Model) are as described in the following:

The first step towards implementing our prediction is stated to the gathering of data sets. Then the second step is processing the Data's we have collected. The third step follows as splitting the data's as we are going to run the Machine Learning Algorithms for getting our desired outputs. Upon pre-processing, every classifier is being used to categorize the pre-processed data. On the fourth step we are going to evaluate the performance. After that patient details will be given to predict the Cardiovascular Disease. Finally, we implemented the proposed model and assessed it for accuracy and performance using several performance indicators. Using various classifiers, an efficient system for predicting heart disease has been created in this model.

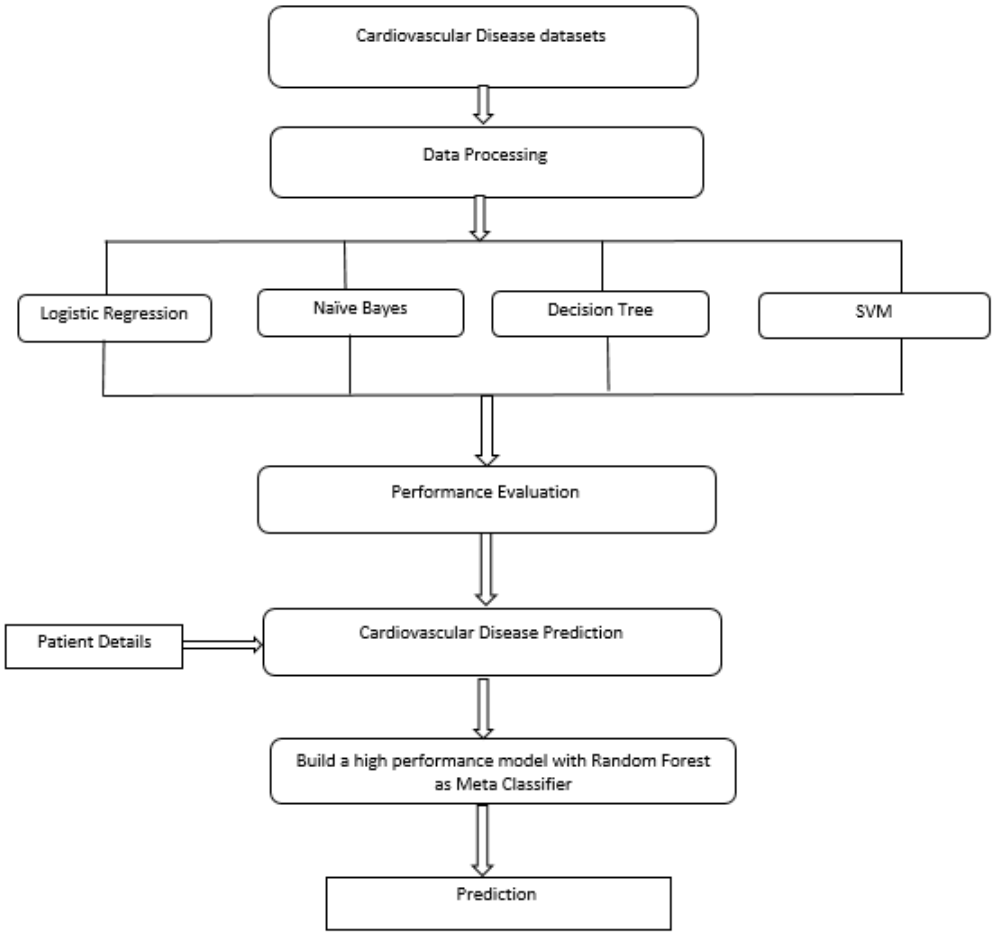


Figure 5.4: Cardiovascular Disease Prediction Model

5.4 Use of the necessary software in a justified manner

The dataset and the Jupyter notebook editor, which employed Python's package manager, were necessities for this project. Jupyter Notebook was the first web application for creating and sharing computational documents. A computational notebook, or in our case Jupyter notebook, is an open source, interactive and free online component. It provides a straightforward, efficient, document-focused experience. The researchers can mix software code, numerical performance, explanatory language, and multimedia technologies into a single document. Through the use of this app, anyone can view and share documents that contain images, live code, computations, visualizations, narrative writing, data cleansing and transformation, numerical simulation, mathematical modeling, data visualization, machine learning, and other features.

For applications in data science that use Python, there is a platform called Jupyter.In addition, it would make working as a data scientist easier because Documentation, data visualization, and caching have all been greatly simplified using Jupyter. It caches the output of each running cell, regardless of whether the algorithm is downloading gigabytes of data from a distant server or training an ML model. Programming environments that are independent of platforms and languages include Jupyter Notebook.

Many languages can be used to understand Jupyter. It also involves the display of some datasets' images and charts, which are produced using scripts and modules. The more Jupyter is being used, the easier it is for developers to describe their codes by line with insights attached. Once the code is fully functional, further explanations and interaction may be added.

Chapter 6

Working Plan/ Methodology

6.1 Importing necessary libraries

Necessary libraries are imported in Jupyter notebook for our thesis project. Some of the libraries are mentioned below with their proper function and how they helped us achieve our goal.

For data analysis and splitting we are using the following libraries:

Pandas: It is a free machine learning package that offers a selection of analysis tools and configurable high-level data structures. It makes data management, cleaning, and analysis easier. Sorting, re-indexing, iteration, concatenation, data conversion, visualizations, aggregates, and other operations are supported by Pandas.

Numpy: Python's numerical programming language is known as Numpy. It is a well-known machine learning library that works with big matrices and multidimensional data. It has built-in mathematical functions that make computations simple.

Scikit-learn: Scikit-learn is a well-known Python toolkit for handling complicated data.

For data visualization we are using the following libraries:

Matplotlib: Plotting numerical data is handled by the package Matplotlib. It is utilized in data analysis because of this. It plots highly specified images like pie charts, histograms, scatterplots, graphs, etc. It is also an open-source library.

Seaborn: Python has a module called Seaborn for creating statistical visuals. It is based on matplotlib and tightly integrated with pandas data structures.

For machine learning we are using the following libraries:

Scikit-learn: Machine learning is supported by the open-source software Scikit-learn. It is compatible with a variety of supervised and unsupervised methods, including linear regression, classification, clustering, etc.The library functions with Numpy and Scipy.

6.2 Selection of Features

The properties of any dataset that are used for analysis and prediction are known as the elements of a dataset [9] . The selection of elements is essential to the machine learning process since the performance of the machine learning classifier might occasionally be impacted by irrelevant features. The model execution time is decreased and classification accuracy is increased using proper elements selection.

The first step in the machine learning process is the pre-processing of the data, which is followed by feature selection based on the categorization and evaluation of the modeling performance, and results with increased accuracy. For different combinations of attributes, the feature selection and modeling process is continually repeated. The effectiveness of each model created using Machine Learning methods are employed for each iteration based on 14 features and recording both the performance.

The proper explanation of Features which we are using on our data set is given in the below table:

Serial N	Features	Description	Values
1	Age	Age of patient	Age in years
2	Sex	Sex of patient	1: male, 0: female
3	Cp	Chest pain	0: Typical angina, 1: Atypical angina, 2: Non-angina pain, 3: Asymptomatic.
4	Trestbps	Resting Blood Pressure	Blood pressure at rest (mm), 120/80 is the usual range of normal blood pressure reading
5	Chol	Serum Cholesterol	Measured by serum cholesterol (mg/dL), it ought to be lower than 170 mg/dL
6	Fbs	Fasting Blood Sugar	>120 mg/dl of fasting blood sugar (1 true), normal blood sugar levels are less than 100 mg/dL (5.6 mmol/L)
7	Restecg	Resting Electrocardiograogic Results	Resting electrocardiographic results (values 0,1,2)
8	Thalach	Maximum Heart Rate	Maximum Heart Rate Achieved by patient
9	Exang	Exercise Induced Agina	Angina induced by exercise (1 yes)
10	Oldpeak	ST depression induced by exercise relative to rest	ST depression induced by exercise relative to rest (values 0 - 6.2)
11	Slope	The slope of the peak exercise ST segemnt	ST segment measured in terms of slope during peak exercise, 0: Unsloping, 1: Flat, 2: Downsloping
12	Ca	Number of major vessels colored by fluoroscopy	Number of major vessels (0–3) colored by fluoroscopy.
13	Thal	Thalassemia	3: normal, 6: fixed defects, 7: reversible defects
14	Target	Angiographic disease status	0: No disease, 1: Disease

Figure 6.1: Features Selection

6.3 Description of features:

Following are the information about the features used in the data set:

```
age:                    age
sex:                    1: male, 0: female
cp:                     chest pain type, 0: typical angina, 1: atypical angina, 2: non-anginal pain, 3: asymptomatic
trestbps:                       resting blood pressure
chol:                    serum cholestoral in mg/dl
fbs:                    fasting blood sugar > 120 mg/dl
restecg:                        resting electrocardiographic results (values 0,1,2)
thalach:                         maximum heart rate achieved
exang:                  exercise induced angina
oldpeak:                        oldpeak = ST depression induced by exercise relative to rest
slope:                  the slope of the peak exercise ST segment
ca:                     number of major vessels (0-3) colored by flourosopy
thal:                   3 = normal; 6 = fixed defect; 7 = reversable defect
target:                 0: No Disease, 1: Disease
```

Figure 6.2: Description of Features

The description of the features include all the features that are included in the data set. Different features have different ranges and contributes differently to the occurrence of cardiovascular disease. For example elderly people are more likely to suffer from cardiovascular disease also people with high cholesterol might have a high chance of developing cardiovascular disease etc.

6.4 Exploratory Data Analysis

Several indicators, including age, gender, pulse rate, cholesterol and others can be used to determine the presence of heart disease. Data analysis in healthcare helps with disease prediction, better diagnosis, symptom analysis, providing suitable medications, enhancing treatment quality, lowering costs, prolonging patient lifespan, and lowering the death rate among cardiac patients. The features below are taken from our data set for better analysis and understanding of our work.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0

Inference: The Datset consists of 14 features & 1025 samples.

Figure 6.3: Features

6.5 Frequency of Features

Age (age) Frequency: Age is a very crucial factor when it comes to the detection of cardiovascular disease. In this paper we are trying to explain how 'age' plays a very crucial role in developing cardiovascular disease. In the diagram below we can see the frequency of people in terms of age distribution. We can see the most frequent age in our data set is 58 with the target variable 0.

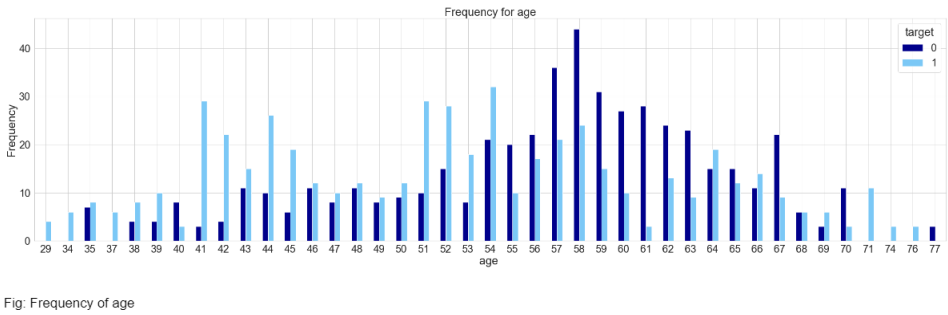


Figure 6.4: Frequency of Age

Gender (Sex) Frequency: Another vital feature in our data set is 'sex' which is correlated to cardiovascular disease. Here is the distribution of 'sex' along with the target variable to understand this feature better and how it is correlated with cardiovascular disease. The following diagram shows the frequency of 'sex' in our data set. We can see gender distribution of male is more frequent with target variable value 0.

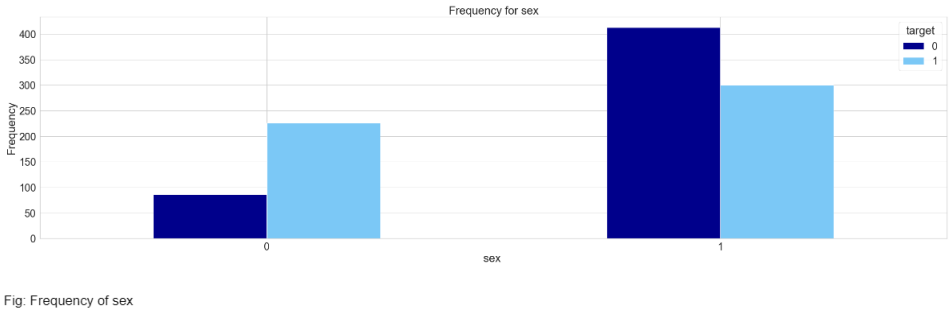


Figure 6.5: Frequency for Sex

Chest Pain (cp) Frequency: 'Chest pain', which is linked to cardiovascular disease, is an essential element of our data set. To better comprehend this characteristic and its relationship to cardiovascular disease, the distribution of 'chest pain' and the target variable are shown here. The frequency of 'chest pain' in our data set is represented in the following figure. We can see 'chest pain' is most frequent at the value 0 with target variable value 0.

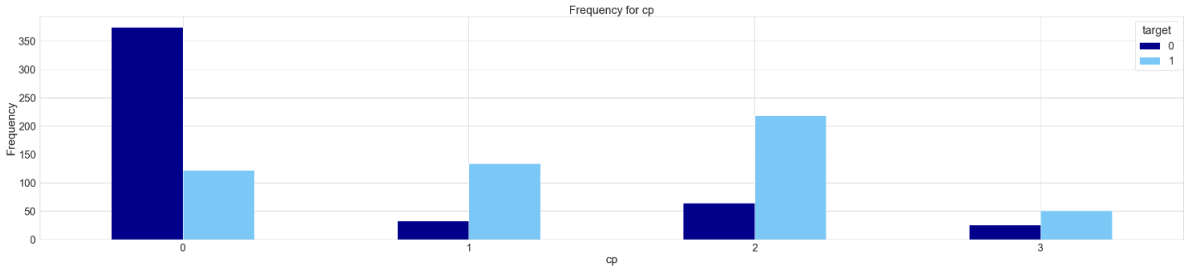


Fig: Chest Pain Frequency

Figure 6.6: Frequency for Chest Pain

Resting blood pressure (trestbps) Frequency: 'Resting blood pressure' is a crucial component of our data set which is closely associated to cardiovascular disease. The distribution of 'resting blood pressure' and the target variable are shown here to help the reader comprehend its correlation with cardiovascular disease. The following figure shows how frequently 'resting blood pressure' occurs in our data set. We can see that 'resting blood pressure' is most frequent at value 130 with the target variable value 1 is most frequent.

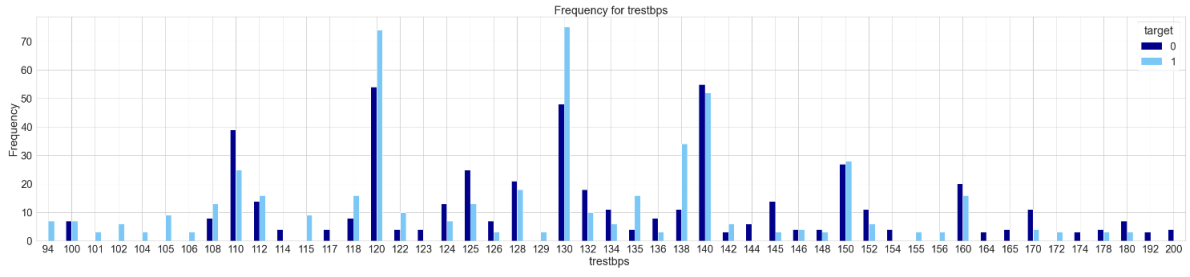


Fig: Frequency of trestbps

Figure 6.7: Frequency of trestbps

Serum Cholesterol (chol) Frequency: The following figure shows the frequency of 'serum cholesterol' in our data. It is very evident that 'serum cholesterol' frequency is highest at 130 with a target value of 1. That means that this is the highest occurring value in our data for 'serum cholesterol'.

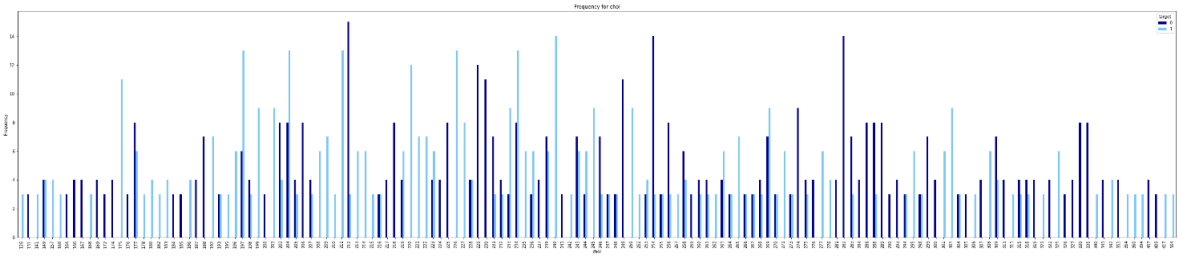


Fig: Frequency of chol

Figure 6.8: Frequency of chol

Resting Electrocardiographic Results (Restecg) Frequency: The reader can better understand the target variable's association with cardiovascular disease by studying the distribution of 'Resting Electrocardiographic Results'. The frequency of "Electrocardiographic Results" in our data set is depicted in the following figure. We can observe that 'Resting Electrocardiographic Results' is occurring most frequently at value 1 with the target variable set to 1.

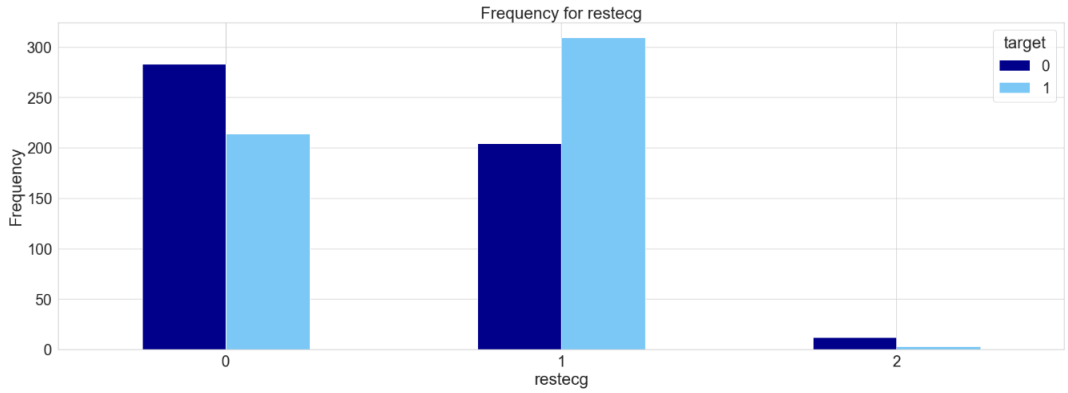


Fig: Frequency of restecg

Figure 6.9: Frequency of restecg

Resting Exercise Induced Angina (exang) Frequency: The term "Exercise Induced Angina" is significant to our data set because it is very closely connected to cardiovascular disease. The target variable and the distribution of "Exercise Induced Angina" are shown below to help understand this trait and its connection with cardiovascular disease. With the target variable set to 0, we can observe that "Exercise Induced Angina" occurs most frequently at the value of 0.

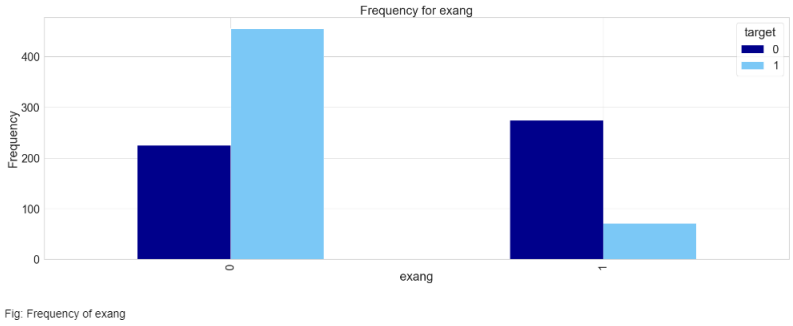


Figure 6.10: Frequency of exang

Maximum Heart Rate Acheived (thalach) Frequency: 'Maximum Heart Rate Achieved', which is associated to cardiovascular disease, is a crucial component of our data set. To better comprehend this feature and how it is connected with cardiovascular disease, the distribution of 'Maximum Heart Rate Achieved' is shown here together with the target variable. The frequency of 'Maximum Heart Rate Achieved' in our data set is illustrated in the following diagram. With the target variable's value of 1, we can determine that the most frequent 'Maximum Heart Rate Achieved' value is 162.

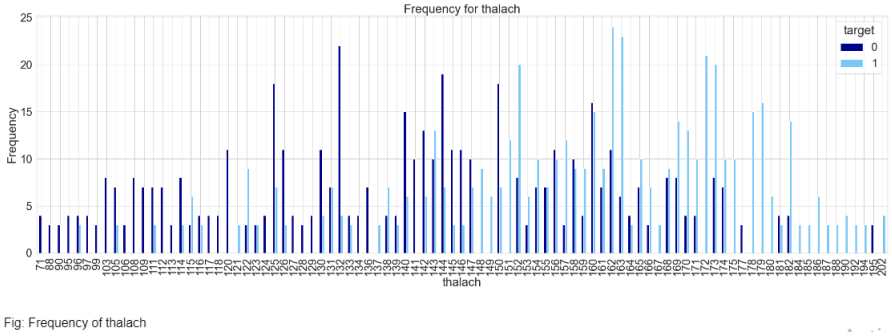


Figure 6.11: Frequency of thalach

ST depression induced by exercise relative to rest (oldpeak) Frequency: The frequency of 'ST depression induced by exercise relative to rest' in our data is shown in the following figure. It is clearly visible that the frequency of 'ST depression induced by exercise relative to rest' peaks at 0.0 with a target value of 1. That indicates that 'ST depression induced by exercise relative to rest' is most frequent at this point.

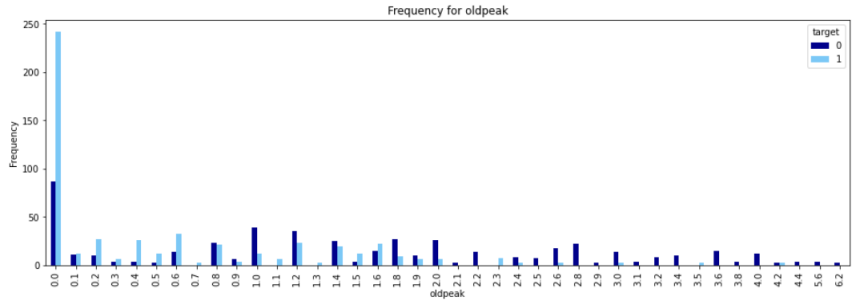


Fig: Frequency of oldpeak

Figure 6.12: Frequency of oldpeak

Slope of the peak exercise ST segment (slope) Frequency: The following diagram illustrates the frequency of the "Slope of the Peak Exercise ST Segment" in our data. It is evident that, with a target value of 0, the frequency of the "Slope of the peak exercise ST segment" peaks at 1. This suggests us that at this point, the "Slope of the Peak Exercise ST Segment" is most prevalent.

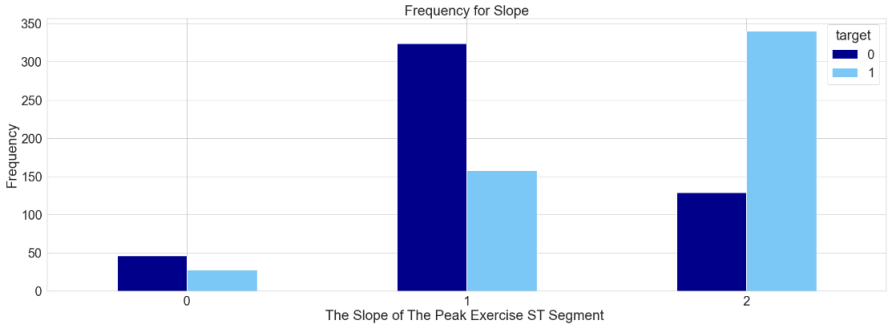


Fig: Frequency of slope

Figure 6.13: Frequency of Slope

6.6 Analysis of Features

Analyzing 'age' feature: Age is a very significant feature that helps us understand any disease better. Now we are trying to understand how cardiovascular disease is related to the age factor with our analysis. The diagrams below gives us a better understanding about the correlation between age and the other variables that helps us predict the model more accurately and efficiently. The first figure illustrates the target-age distribution diagram. The second figure illustrates density-age distribution diagram. And the third diagram represents thalach-age scatter plot diagram, which clearly indicates if a person has the disease or not.

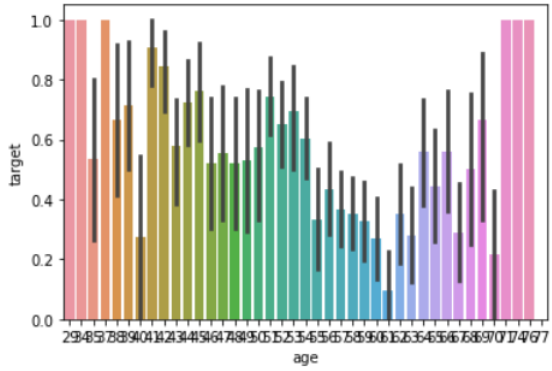


Fig: Target-Age Feature Diagram

Figure 6.14: Target-Age-Feature

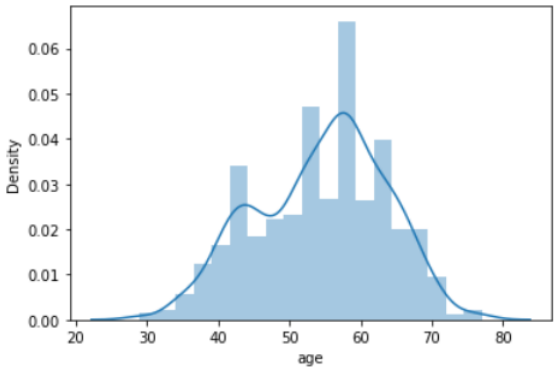


Fig: Density-Age Diagram

Figure 6.15: Density-Age

From the scatter plot diagram below, it evident that people in between the age of 40 to 60 has a higher rate of maximum heart rate which is a clear indication of cardiovascular disease. Therefore it can be said that people in between this age limit are more likely to develop cardiovascular disease.

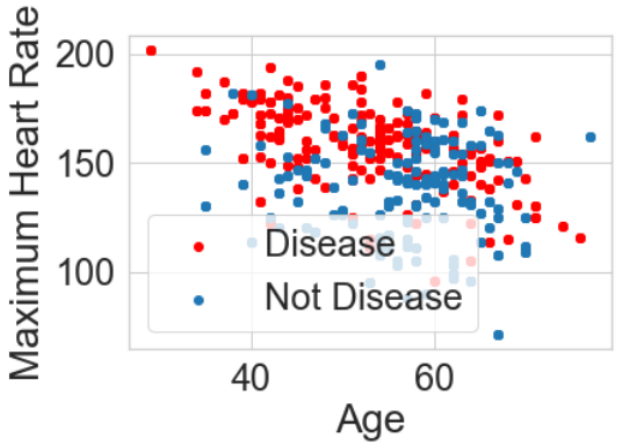


Fig: Thalach-Age Scatter plot diagram

Figure 6.16: Thalach-Age Scatter Plot Diagram

Analyzing 'cp' feature: Chest pain is a very important attribute that contributes in our understanding of cardiovascular disease. With our data, we are currently attempting to comprehend how cardiovascular illness is connected to the cp element. The diagrams below help us better grasp the relationship between cp and the other variables, which improves the effectiveness and precision of our model prediction. The target-cp distribution diagram is depicted in the first figure. The density-cp distribution diagram is depicted in the second figure. The final diagram is a thalach-cp scatter plot diagram, which makes it evident whether or not a person has the condition. It is very evident from the diagram that cp type 1 has the highest possibility of having cardiovascular disease.

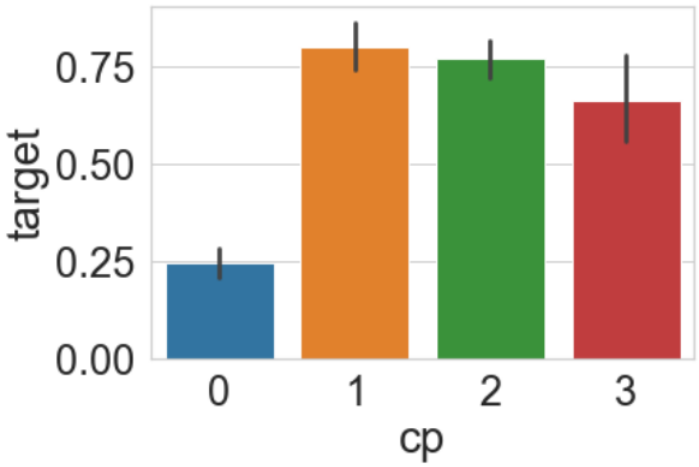


Fig: Target-cp feature diagram

Figure 6.17: Target-cp Feature Diagram

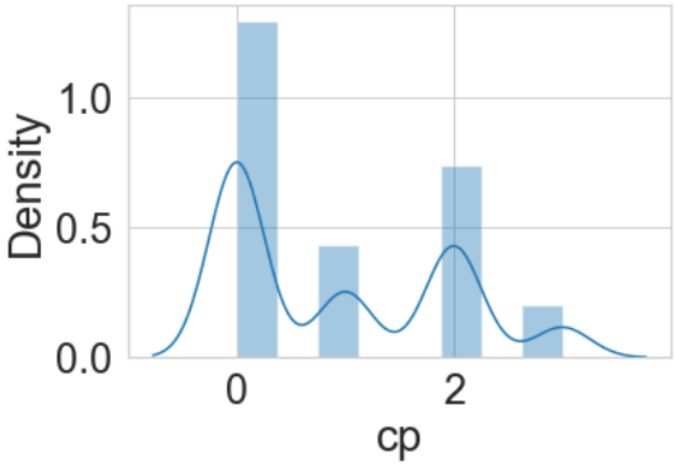


Fig: Density-cp Diagram

Figure 6.18: Density-cp Diagram

In the figure below it is very clear that the rate cardiovascular disease is highest when the value of cp is 1. The scatter plot diagram also demonstrates that cardiovascular disease is more likely to occur when cp is of type 1, 2 and 3. It is safe to say that the rate of cardiovascular disease is least when cp is of type 0.

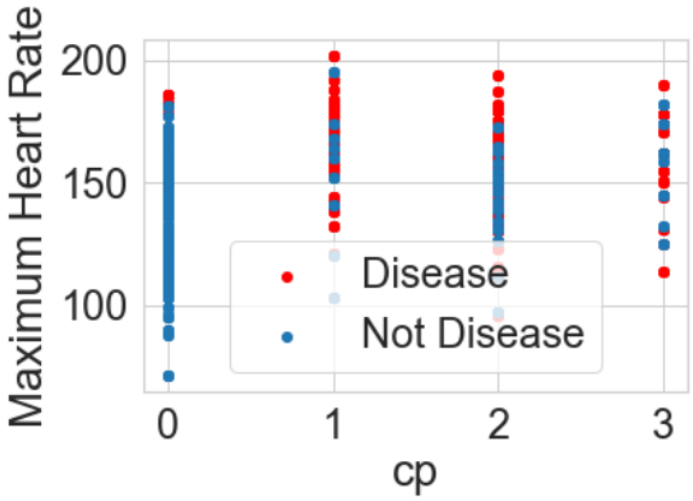


Fig: thalach-cp scatter plot Diagram

Figure 6.19: Thalach-cp Scatter Plot Diagram

Analyzing 'fbs' feature: Fasting blood pressure is a significant component that assists us evaluate cardiovascular disease and the factors associated with it. In our research, we are currently seeking to appreciate how cardiovascular sickness is related to the fbs aspect. The relationships between fbs and the other variables are easier to understand thanks to the diagrams below, which also increases the effectiveness and accuracy of our model's prediction. The first figure shows the target-fbs distribution diagram. The second figure shows the density-fbs distribution diagram. A thalach-fbs scatter plot diagram in the end diagram shows whether or not a person has the condition.

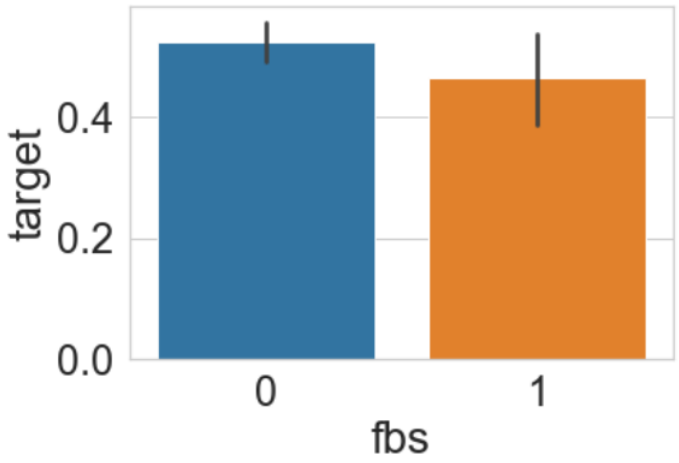


Fig: Targer-fbs feature Diagram

Figure 6.20: Target-fbs Feature Diagram

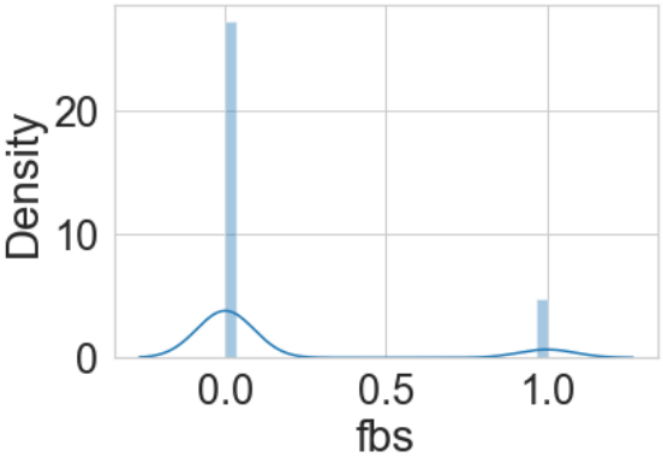


Fig: Density-Fbs Diagram

Figure 6.21: Density-fbs Diagram

In the figure below the relationship between maximum heart rate achieved and fasting blood pressure (fbs) is clearly demonstrated. We can understand how they are correlated and how change in one feature can result in a significant change in the other one. It is very clear that a person of fbs value 1 is more likely to develop cardiovascular disease.

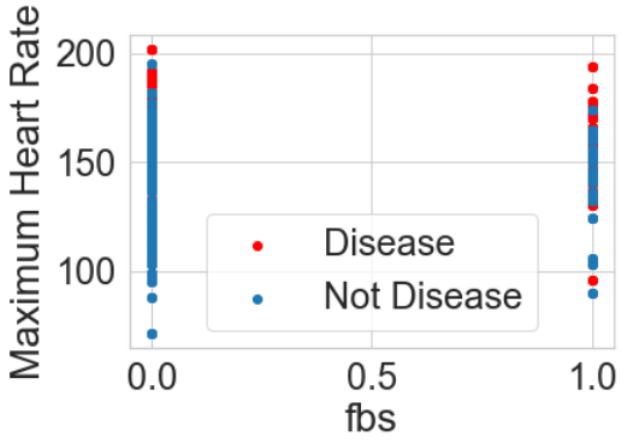


Fig: thalach-fbs scatter plot diagram

Figure 6.22: Thalach-fbs Scatter Plot Diagram

Analyzing 'restecg' feature: Resting Electrographic Results play a big role in how we assess cardiovascular disease and the risk factors that contribute to it. We are currently trying to understand how cardiovascular disease relates to the restecg aspect in our research. The graphs below make it simpler to comprehend the links between restecg and the other variables, which also improves the predictive power and precision of our model. The target-restecg distribution diagram is shown in the first figure. The density-restecg distribution diagram is shown in the second figure. The final diagram displays a thalach-restecg scatter plot diagram that indicates whether or not an individual has the condition.

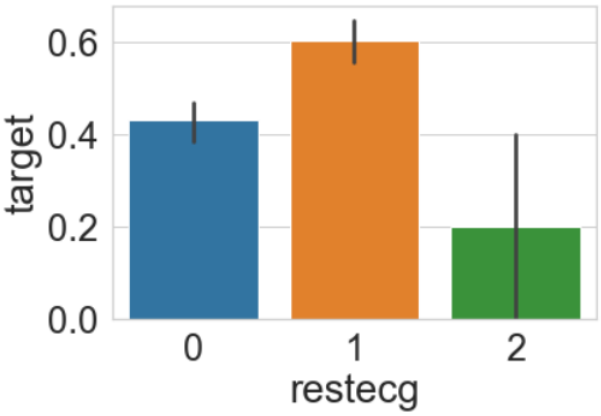


Fig: Target-restecg feature Diagram

Figure 6.23: Target-regtecg Feature Diagram

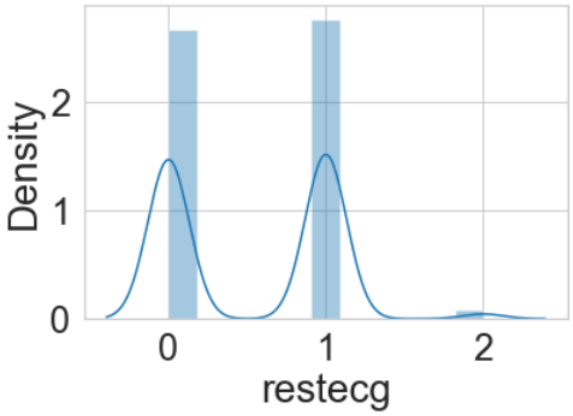


Fig: Density-restecg Diagram

Figure 6.24: Density-restecg Diagram

The rate of cardiovascular disease is at its maximum when the value of restecg is 0, as is shown in the figure below. The scatter plot diagram also illustrates that when the value of restecg is 0 and 1 there is an increased risk of cardiovascular disease. It is safe to assume that restecg value 2 has the lowest rate of cardiovascular disease.

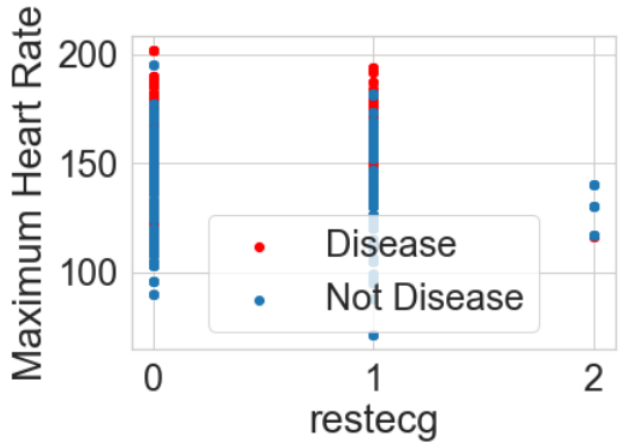


Fig: thalach-restecg scatter plot Diagram

Figure 6.25: thalach-restecg Scatter Plot Diagram

Analyzing 'sex' feature: When evaluating cardiovascular disease and the risk factors that lead to it, gender plays a significant influence. In our research, we are currently attempting to comprehend how sex is related to cardiovascular disease. The relationships between sex and the other factors are easier to understand due to the graphs below, which also enhances the accuracy, reliability and predictive effectiveness. In the first figure, the target-sex distribution diagram is represented. The second figure displays the density-sex distribution diagram. The thalach-sex scatter plot diagram in the final figure illustrates whether or not a person has the disease.

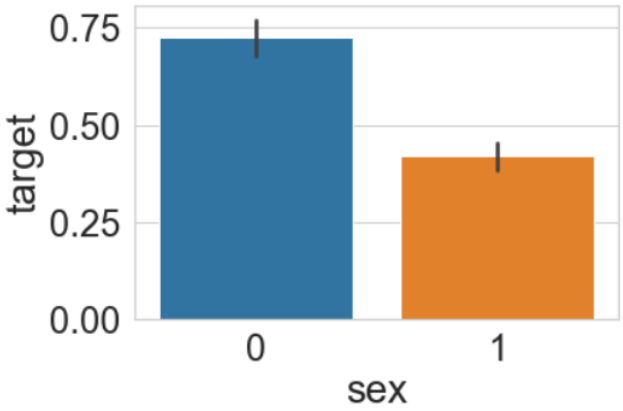


Figure 6.26: Target-sex Diagram

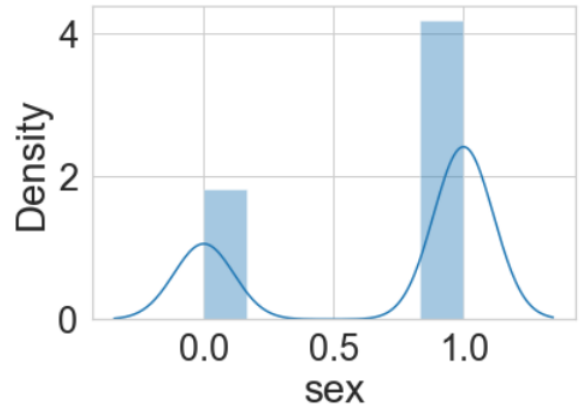


Fig: Density-Sex Diagram

Figure 6.27: Density-sex Diagram

In the figure below it is evident that maximum heart rate achieved is correlated to the gender of the person and it plays an important role in determining and indicating if a person has cardiovascular disease or not.

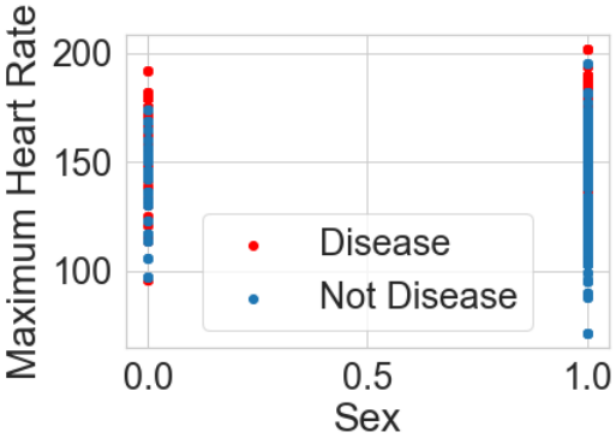


Fig: thalach-sex scatter plot Diagram

Figure 6.28: thalach-sex Scatter Plot Diagram

Analyzing 'exang' feature: Now let's examine the 'exang' feature to gain a better understanding of it. Exercise Induced Angina is a major element to consider when assessing cardiovascular disease and the risk factors that contribute to it. We are actively investigating the relationship between exang and cardiovascular illness in our research. The graphs below make it simpler to grasp the relationships between exang and other parameters, which also improves the accuracy, dependability, and prediction efficacy. The target-exang distribution diagram is shown in the first figure. The density-exang distribution diagram is shown in the second figure. The last figure shows a thalach-exang scatter plot diagram that shows whether or not a person has the condition.

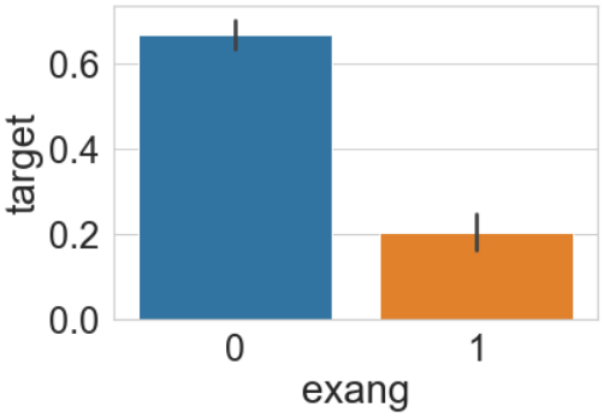


Fig: Target-Exang feature diagram

Figure 6.29: Target-exang Feature Diagram

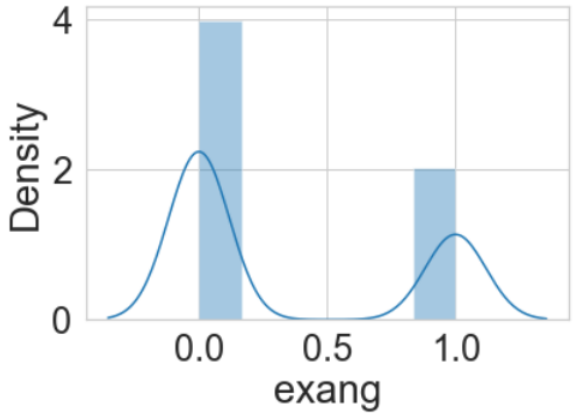
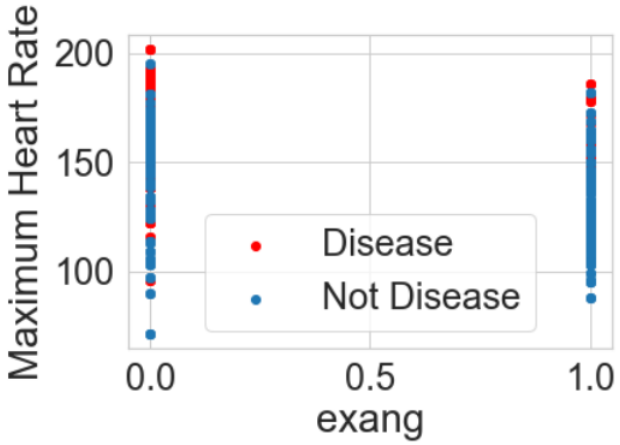


Fig: Density-Exang feature Diagram

Figure 6.30: Density-exang Feature Diagram

The association between maximum heart rate achieved and exercise-induced angina (exang) is readily demonstrated in the figure below. We understand how closely linked they are and how altering one of them may have a big impact on the other. It is pretty apparent that someone who has an exang value of 0 has a higher risk of developing cardiovascular disease.



thalach-exang scatter plot diagram

Figure 6.31: thalach-exang Scatter Plot Diagram

Analyzing 'target' feature: The target feature in our data set is a very important feature. It determines whether a person has cardiovascular disease or not. This feature is measure of predicting cardiovascular disease in our model. Analyzing this feature is vital to understand the disease better. After analyzing this feature it is very evident that this feature is very strongly correlated with maximum heart rate achieved (thalach), which is an important indicator of cardiovascular disease prediction. It is very also very evident from the diagram that when the target value is one, a person is very likely to have cardiovascular disease.

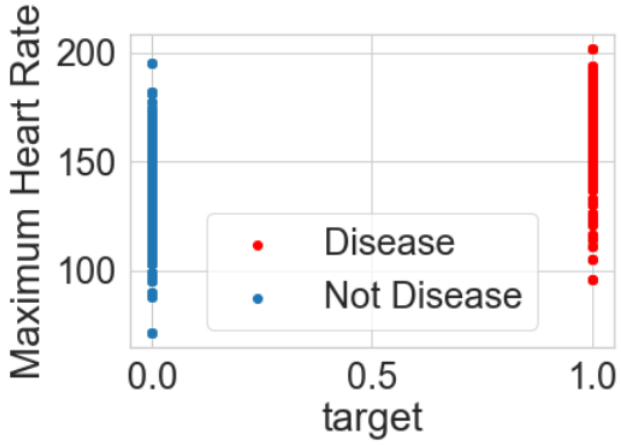


Fig: thalach-target scatter plot Diagram

Figure 6.32: thalach-target Scatter Plot Diagram

6.7 Correlation Between Attributes

The correlation between various cardiovascular disease features is displayed in the following diagram. It is apparent from the figure below that these characteristics are linked to cardiovascular disease and can serve as a marker for the disease. These factors' distribution exhibits a normal distribution. There is a positive correlation between some of these attributes which means these features play a vital during the development of this disease.

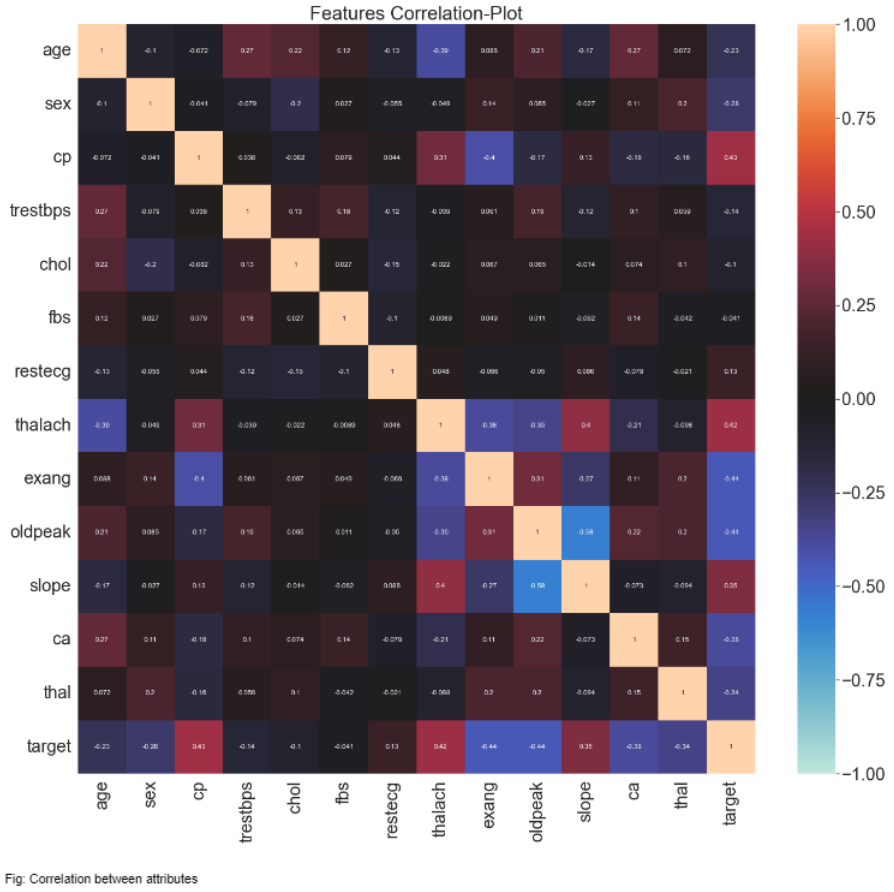


Figure 6.33: Correlation Between Attributes

From the figure above we can come into the conclusion that, 'cp, 'thalach' and 'slope' exhibit a strong positive correlation with the target, which means these features are strongly linked with occurrence of cardiovascular disease. 'Age', 'exang', 'ca', 'thal', 'gender', and 'oldpeak' all have strong negative correlations to the target. 'Fbs', 'chol', 'trestbps' and 'restecg' shows a low correlation with the target.

The following diagram shows the correlation of different features with our target variable. The relationship between the individual feature and the target variable is shown in the diagram. This diagram helps us understand the relationship between these attributes with the target variable which helps us understand the disease even better.

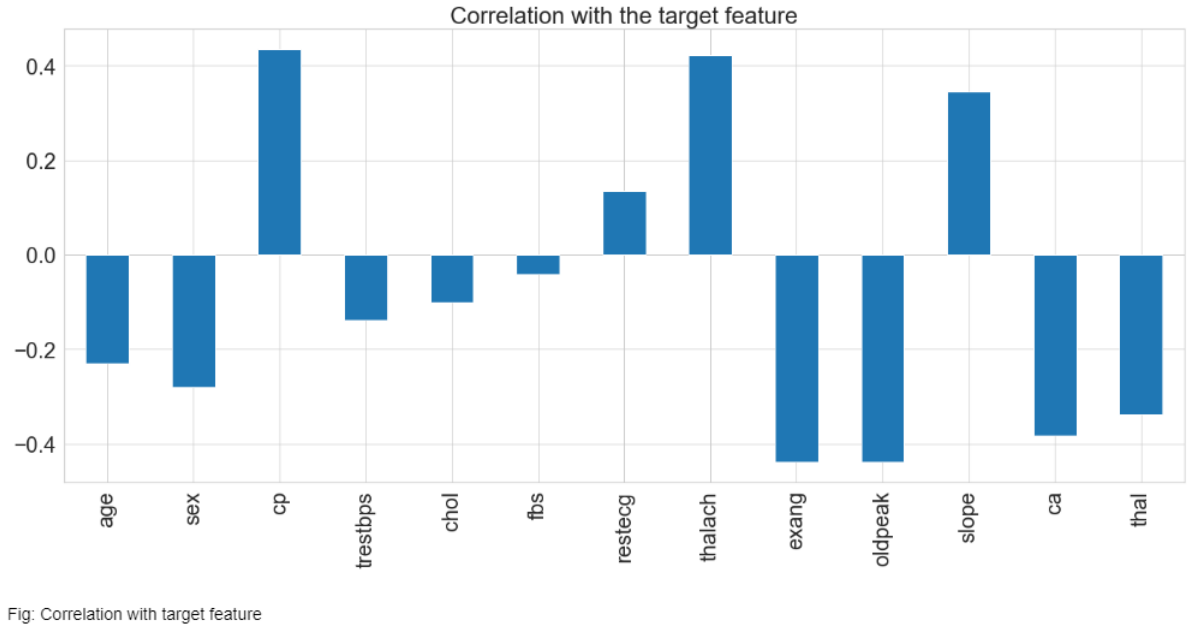


Figure 6.34: Correlation with Target Feature

The figure demonstrates the correlation between features in our dataset and the target variable. It is clearly illustrated that the features 'cp', 'restecg', 'thalach' and 'slope' are all positively correlated with the target variable which means they are very closely linked to cardiovascular disease. They are very strongly related to our target feature and changes in them can have an impact in our target variable. Other features in the diagram shows negative correlation with our target variable.

6.8 Demonstration of Healthy and Cardiovascular Disease patients

Cardiovascular diseases are generally referred to as heart and blood vessel diseases. Heart attacks and strokes are typically sudden, catastrophic occurrences that are mostly caused by a blockage that stops the flow of blood to the heart or brain. Fatty deposits that have developed on the inner walls of the blood arteries that connect the heart or brain are the most frequent cause of this. Blood clots or hemorrhage from a brain blood artery can both result in strokes. One or more areas of your heart and/or blood vessels may be impacted by these diseases. A person may exhibit symptoms of the disease, physical manifestations or not feeling anything at all. The following diagram illustrates the percentage of healthy and cardiovascular disease patients present in our data set.

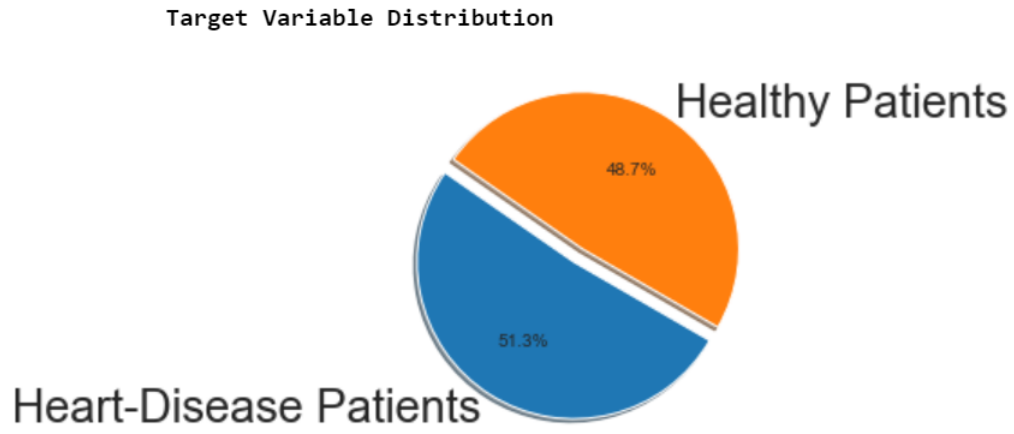


Fig: Distribution of Cardiovascular disease patients in our data

Figure 6.35: Distribution of Cardiovascular Disease patients in our Data

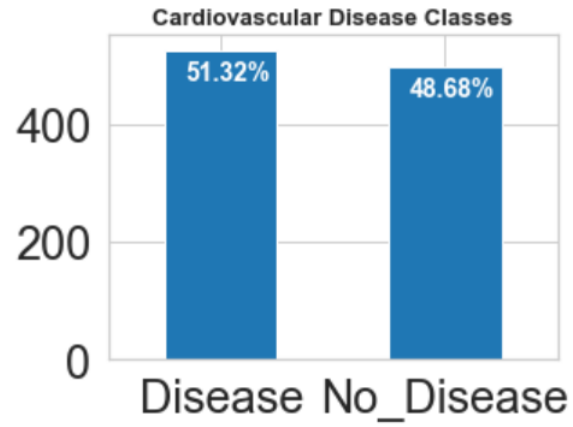


Fig: Cardiovascular disease classes

Figure 6.36: Cardiovascular Disease Classes

After carefully examining and analyzing our data we figured out that there are 51.32% patients with cardiovascular disease and the rest 48.68% people are healthy. Taking various features into consideration we have come into this conclusion. Therefore it is very evident that we have more cardiovascular disease patients than healthy people in our data.

6.9 Feature Distribution

Feature distribution is the distribution of a feature over its range, with value on the horizontal axis and frequency on the vertical axis. By segmenting the range into a number of bins and figuring out how many data points fall within each bin's limits, the feature distribution is displayed as a histogram. Understanding the type of feature we are dealing with and the values we can anticipate from it are made easier by looking at the feature distribution. We'll examine to verify if the values are evenly distributed or not. Now let us take a look at the distribution of some of the important attributes of our research. The following diagram illustrates the feature distribution of our data in this research.

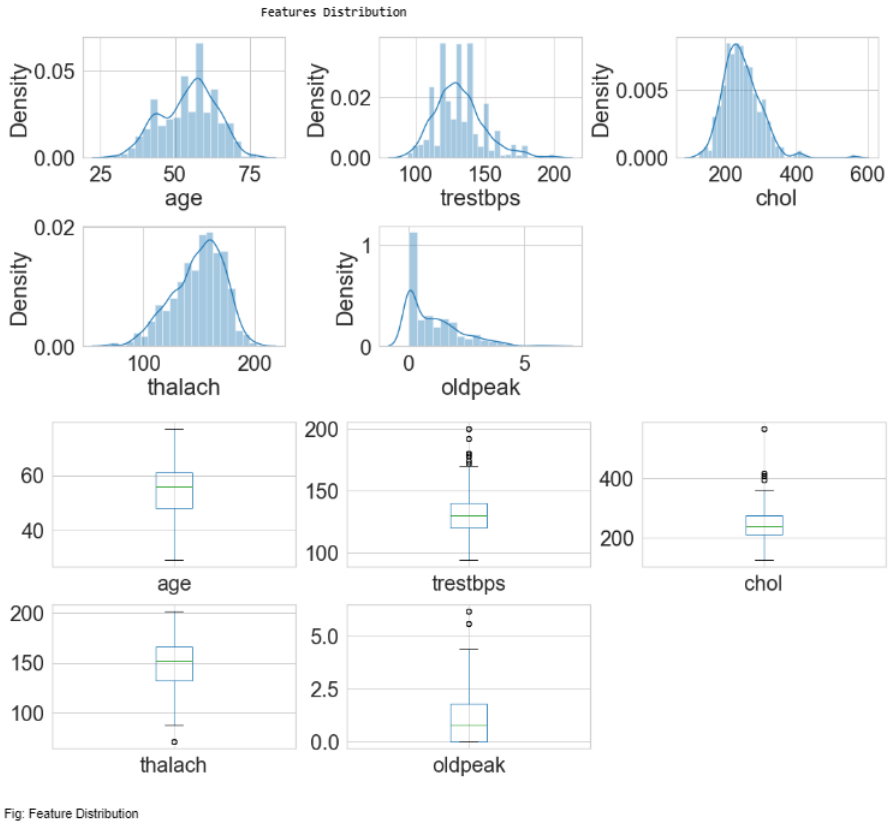


Fig: Feature Distribution

Figure 6.37: Features Distribution

6.10 Feature Engineering

Feature engineering is the process of selecting, manipulating, and transforming raw data into features that may be used in supervised learning. To make machine learning effective on new tasks, it may be necessary to develop and train better features. Any quantifiable input that can be employed in a predictive model is known as a feature. Feature engineering, in its simplest form, is the process of using statistical or machine learning approaches to convert raw data into desired features. Our data after performing feature engineering is demonstrated below.

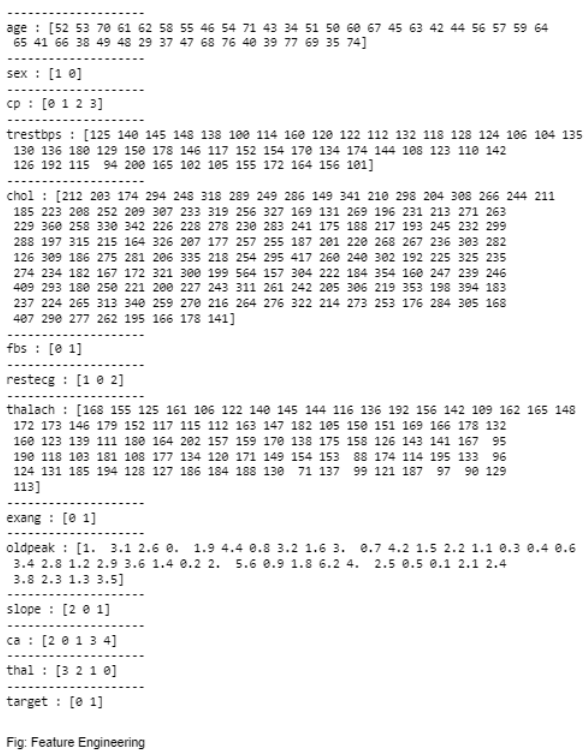


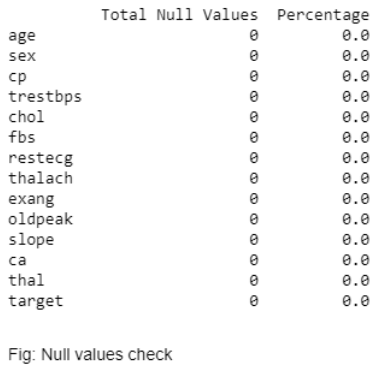
Figure 6.38: Feature Engineering

6.11 Data Preprocessing

Data analysis or pre-processing will be essential for our prediction model, as we will not be able to generate reliable results from machine learning algorithms without it. Data must be preprocessed in order for it to be represented effectively and for a machine learning classifier to be trained and tested properly. By transforming medical records into diagnosis values, the pre-processing of data is carried out.

As par our knowledge some of the Machine Learning algorithms does not support missing values. So, in order to resolve such issues, we must manage null values from the original raw data. As a result, two outputs will be generated: one will show patients who have been identified as having no disease, and the others will show patients who have been identified as having cardiovascular diseases.

For the dataset to be used effectively by the classifiers, preprocessing methods such missing value removal, standard scalar can be applied. Simply remove the feature row with missing values from the data set. These data preprocessing methods were employed in this research. To perform data preprocessing first we are checking for empty elements and then we are removing the null variables.



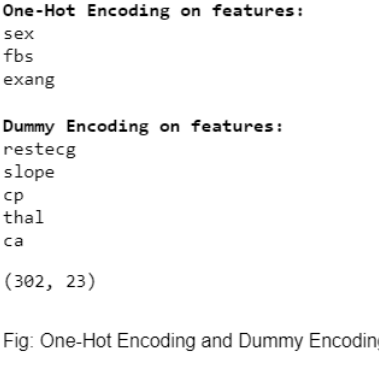
	Total Null Values	Percentage
age	0	0.0
sex	0	0.0
cp	0	0.0
trestbps	0	0.0
chol	0	0.0
fbs	0	0.0
restecg	0	0.0
thalach	0	0.0
exang	0	0.0
oldpeak	0	0.0
slope	0	0.0
ca	0	0.0
thal	0	0.0
target	0	0.0

Fig: Null values check

Figure 6.39: Null Values Check

Then we are performing One-Hot Encoding and Dummy Encoding on features for categorical columns to numeric conversion. One hot encoding method involves transforming categorical information into a format that may be given to ML algorithms to help them perform better at prediction. Dummy (binary) variables are used in dummy encoding. Dummy encoding employs k-1 dummy variables rather than producing a set of dummy variables equal to the number of categories (k) in the variable. After this we are doing the outliner removal to have a processed set of data.

One-Hot Encoding and Dummy Encoding:



One-Hot Encoding on features:
sex
fbs
exang

Dummy Encoding on features:
restecg
slope
cp
thal
ca

(302, 23)

Fig: One-Hot Encoding and Dummy Encoding

Figure 6.40: One-hot Encoding and Dummy Encoding

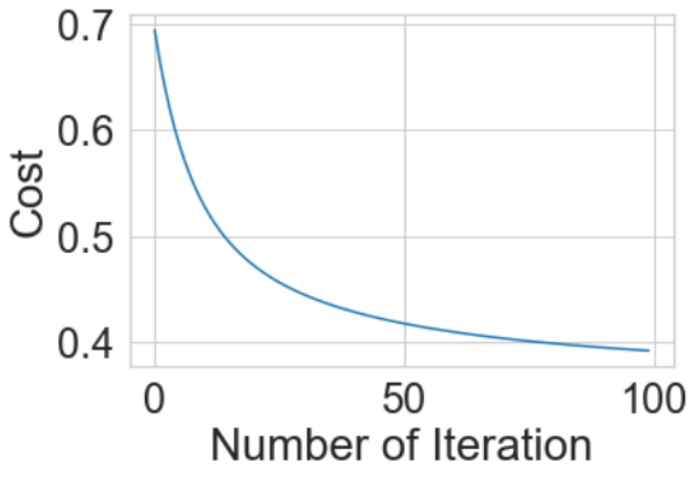
	age	sex	trestbps	chol	fbs	thalach	exang	oldpeak	target	restecg_1	...	cp_1	cp_2	cp_3	thal_1	thal_2	thal_3	ca_1	ca_2	ca_3	ca_4
0	52	1	125	212	0	168	0	1.0	0	1	...	0	0	0	0	0	1	0	1	0	0
1	53	1	140	203	1	155	1	3.1	0	0	...	0	0	0	0	0	1	0	0	0	0
2	70	1	145	174	0	125	1	2.6	0	1	...	0	0	0	0	0	1	0	0	0	0
3	61	1	148	203	0	161	0	0.0	0	1	...	0	0	0	0	0	1	1	0	0	0
4	62	0	138	294	1	106	0	1.9	0	1	...	0	0	0	0	1	0	0	0	1	0

Figure 6.41: Processed Data

Training and Testing: For this research we are training 80% of our data and testing 20% of our data. A portion of our real data set called training data is used by the machine learning model to find and learn patterns. It trains our model in this manner. Training data are typically larger than testing data. This is because we want to provide the model as much data as we can so that it can recognize and pick up on useful patterns. When the information from our data sets is fed into a machine learning algorithm, the program finds patterns and makes predictions.

After building a machine learning model using your training data, we need unknown data to test it. We may evaluate the success and growth of the training of our algorithms using this data, which is referred to as testing data, and then modify or optimize them for better results. When testing data, there are two primary requirements: it must accurately reflect the data set under test and be large enough to enable reliable predictions.

Test Accuracy of Our Data:




```
Test Accuracy: 88.29%
```

Figure 6.42: Test Accuracy of Our Data

Chapter 7

Machine Learning Algorithms for our prediction model

7.1 Logistic Regression

Another supervised ML approach is logistic regression. It associates between dependent and independent variables [9] . It is a supervised learning classification algorithm that can estimate the probability of any given value. This is one of the most basic machine learning algorithm which we are using for predicting cardiovascular diseases. Logistic regression is a machine learning technique that was influenced by the statistical field.

This can be used in classify the binary data which uses two classes to differ between the values. Here, the outcome is predicted using a logistic function, a nonlinear function as opposed to linear regression. The logistic function can changes any number between 0 and 1. This could be crucial if more evidence is needed to support a prediction. This algorithm performs better when qualities related to one another and variables unrelated to the output variable are excluded.

Logistic Regression preferred especially for classification problems involving features from two classes. It is a statistical method that projects future data based on findings from the past using the data set that is currently available. The logistic regression approach predicts a variety of data parameters by looking at the connection between one or more existing predictor factors. The logistic regression equation is depicted in the equation as follows:

$$y = \frac{\left[e^{(\beta 0 + \beta 1 * x + \beta 2)}\right]}{\left[(1 + e^{(\beta 0 + \beta 1 * x + \beta 2)})\right]}$$

After running Logistic Regression in our model here is test accuracy:

Test Accuracy: 85.85%

7.2 Random Forest

Machine learning algorithms like Random Forest are frequently employed to solve data classification and regression issues. The ability of this approach to handle data sets with continuous variables (in the case of regression) and categorical variables is its key strength (in case of classification). Although, for classification problems, Random Forest will be far preferable [4] .

In the case of Random Forest Classifier, we need to have more trees to obtain higher accuracy[9] . In the beginning, it creates call trees using randomly selected knowledge samples from the dataset. The prediction from each tree is then obtained, and the most efficient resolution is chosen using means voting. This is an improvement over decision trees. Image classification, recommendation systems, and feature selection are a few of its applications. Due to the large number of trees working together in this algorithmic rule, it is regarded as a very accurate and robust methodology. The fact that it does not have the over-fitting issue is just one of its many positives. In order to eliminate biases, it then takes the average of all the predictions made by each tree.

It is a machine learning classifier that builds models using Decision Tree Classifiers and is based on supervised learning methods. Trees often identify strange behavior and introduce modest deviations into the training model. It is applied to lessen feature variability in our particular set of data. It also helps with classification on the same training data set. Testing occurs and data sets at the cost of a modest bias increase. Using an ensemble methodology, perform specific tasks such as categorization and future prediction. A class that has been selected by the majority of trees will be generated using Random Forest. Random Forests offer K-fold cross-validation effects.

The decision trees are produced by the random forest supervised machine learning method. To arrive at the final option, the majority of the decision tree is used. Decision trees have problems with low bias and wide variation.

Steps for implementation of Random Forest in our prediction model:

1. Pre-process the data's from the data set.
2. Then, the Random Forest algorithm being modified for the Training set.
3. Then comes prediction of the test result.
4. Make sure the outcome is accurate and also here at this point Confusion matrix will be created.
5. Finally, display the results of the test set.

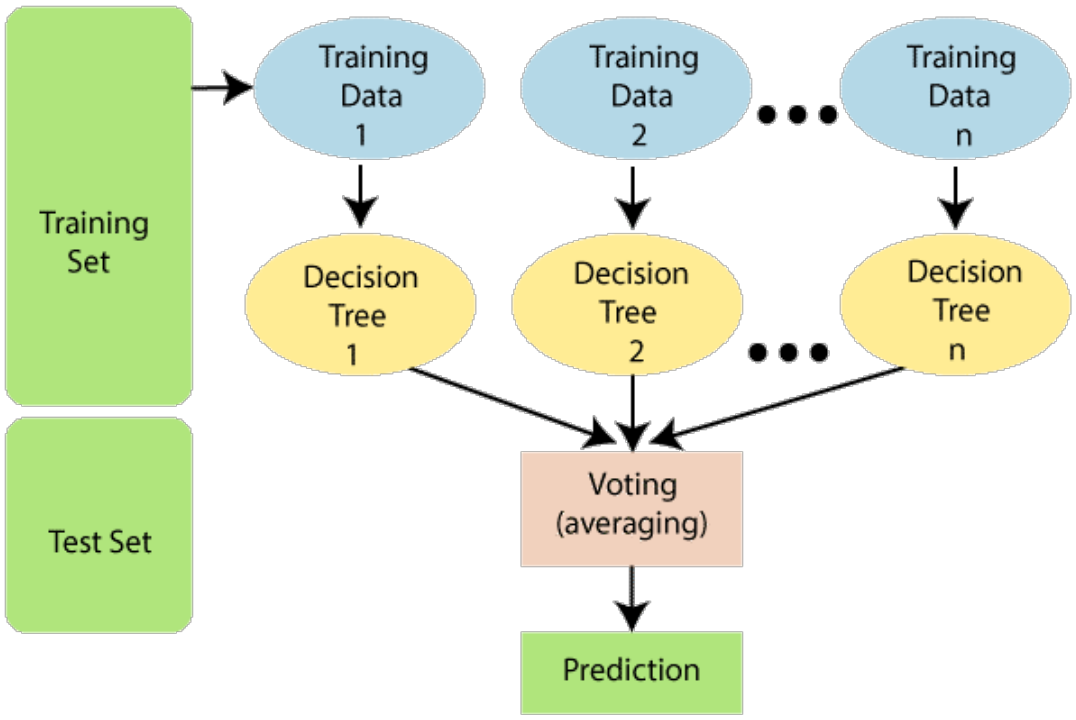


Figure 7.1: How Random Forest Algorithm Works

After running Random Forest in our model here is the test accuracy:

Random Forest Algorithm Accuracy Score: 100.00%

7.3 K-Nearest Neighbors (KNN)

The KNN approach may directly compete with precise existing models due to its high level of accuracy. The KNN technique is ideal if individuals need great precision but do not require a human-readable method. Forecasts can mostly be assessed using distance measures. This technique, which uses "highlight closeness" to foretell the assessments of new data points, assigns a value to a new data point based on how much it resembles the points in the training set.

KNN works in our model in the following steps:

Step 1: Any algorithm needs a data set to be run. As a result, at the very first stage of KNN, we should stack the number of objectives with the test information.

Step 2: The estimation of K must then be chosen, for instance, using the nearest data points. Any whole number can be used as K.

Step 3: Execute the corresponding for each test information point.

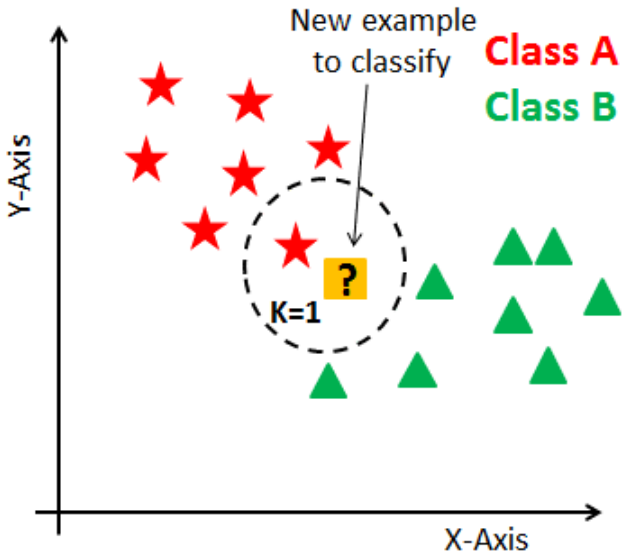


Figure 7.2: How KNN Algorithm Works

After running KNN in our model here is the test accuracy

Maximum KNN Score is: 100.00%

7.4 Support Vector Machine (SVM)

It is a supervised classification-based machine learning (ML) technique that may be used for a variety of regression and classification problems. This is also widely used to eliminate classification and regression tasks.

This approach can be used to handle complex issues that are not amenable to linear solutions. By using a "kernel trick" function, SVM may efficiently find non-linear solutions to any problems. In SVM, the data are projected onto a high-dimensional region to break the challenge into manageable pieces. The classifier chooses the division line with the largest deficit. Every argument is considered as a support.

Support Vector Machines is a classification method that creates hyper planes to divide up input values. Based on the distribution of data, hyper planes can take on a variety of shapes, but only those points that aid in classifying the classes are taken into account. Finding the most extreme minor hyper plane is the main objective of SVM in classifying the data sets.

It works in our model in the following steps:

Step 1: First, SVM will iteratively create hyper planes that separates the classes in the best possible way.

Step 2: The hyper plane that effectively isolates the classes will be chosen at that point.

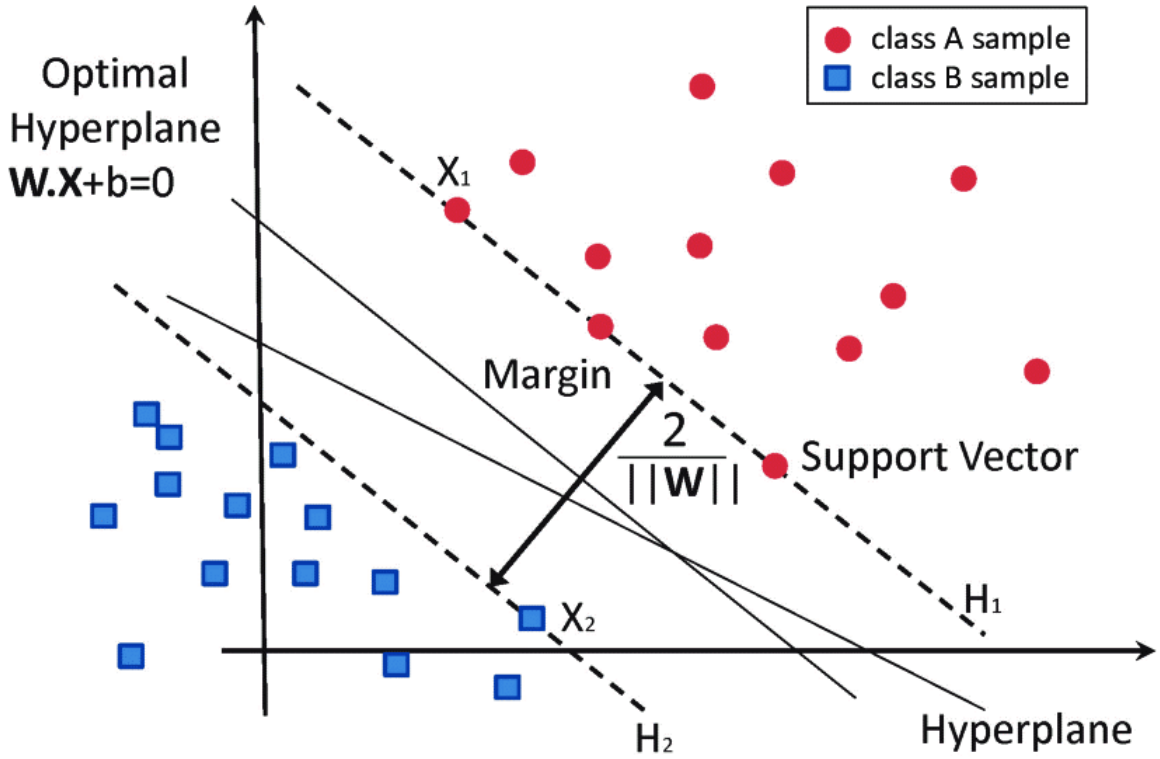


Figure 7.3: How SVM Algorithm Works

The accuracy score achieved after using SVM in our model is:

Test Accuracy of SVM Algorithm: 90.24%

7.5 Decision Tree

A tree like algorithm is basically what DT is. Because of DT's simplicity, a significant amount of data relevant to medicine has been evaluated with its assistance [9] . The trees format is employed since the main goal of the DT technique is to build a model that can predict the value of any target variable. The leaf node corresponds to any class label, and the attributes represent the tree's internal node [25]. Classification and regression issues can be resolved using the supervised learning technique known as a decision tree. On a tree-structured classifier, where each leaf node represents the classification outcome and inside nodes it represent the features of a data set.

Decision Tree works in our model in the following steps:

Step 1: Firstly, we have to start the tree at the root node, because it has the whole data set.

Step 2: Secondly, we are using the Attribute Selection Measure to determine the most important attribute in the data set .

Step 3: Thirdly, create subset to include potential values for the finest attributes.

Step 4: In step 4 we have to create a decision tree which will carry the best attribute.

Step 5: Lastly, Utilize the data set selections from step 3 for periodically create new decision trees. Than the process will continue until we no longer classify the nodes, at that point it will refer to the final node as a leaf node.

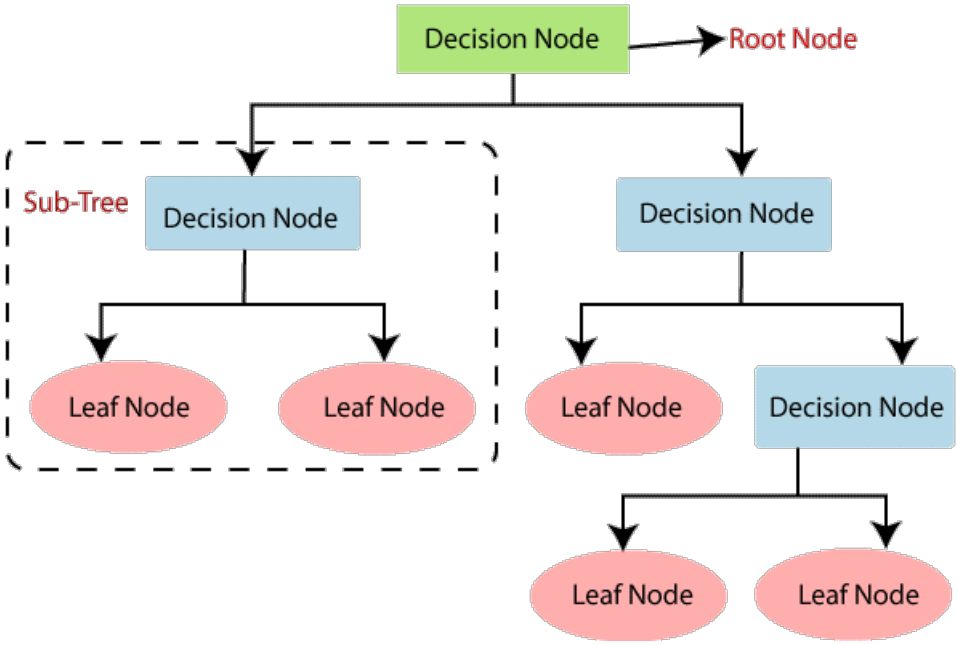


Figure 7.4: How Decision Tree Algorithm Works

The accuracy score achieved after using Decision Tree in our model is:

Decision Tree Test Accuracy 100.00%

7.6 Naïve Bayes

Naive Bayes method's prime focus is to execute the assumption of predictor independence. It is a classifier theorem that can presume that the presence of any specific feature in a class is unrelated to the presence of any other feature, to give a brief definition of what it is. All the parameters present here will work autonomously so that we can get the maximum probability [9].

The training phase is where the Nave Bayes classifier is primarily used. It is based on probability. This approach is used to remove unnecessary information from data sets. Binary classification, also known as two-class and multi-class classification, is based on the fundamental idea of least-squares. The method lends itself to binary classification and variable input data the best.

The Naive Bayes framework is simple and effective for handling vast amounts of data. It yields a higher precision when compared to other machine learning techniques. The Bayes theorem determines the probability that an occurrence will occur based on the possibility of a prior incident.

$$Prob\left(\frac{Ai}{Bi}\right) = \frac{Prob\left(\frac{Bi}{Ai}\right) * Prob(Ai)}{Prob(Bi)}$$

The accuracy score achieved after using Naïve Bayes in our model is:

Accuracy of Naive Bayes: 85.37%

Chapter 8

Results and Discussions

8.1 Accuracy Table

Test Accuracy table of different algorithm used in our model:

Test Accuracy Comparison	
Algorithms	Test Accuracy Percentage
Logistic Regression	85.85%
Random Forest	100.00%
K-Nearest Neighbor (KNN)	100.00%
Support Vector Machine (SVM)	98.24%
Decision Tree	100.00%
Naive Bayes	85.37%

Table 8.1: Test Accuracy Comparison

8.2 Comparing the Accuracy of models

The figure below illustrates the comparison of different models used in our research.

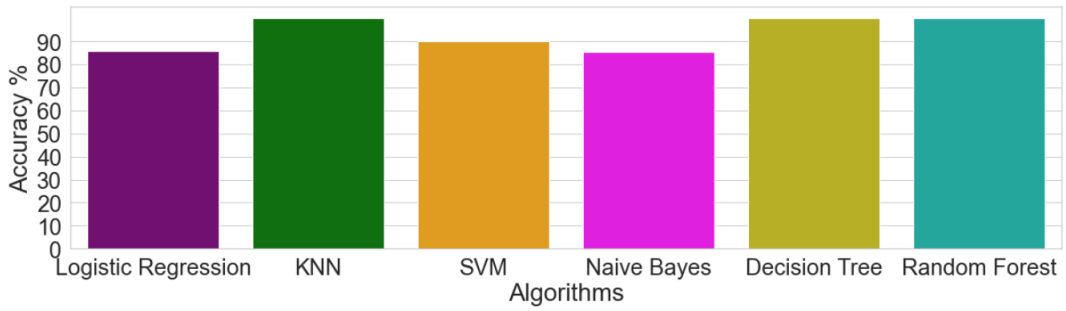


Figure 8.1: Accuracy Graph of Different Models

8.3 Confusion Matrix

The figure below illustrates the confusion matrix of different algorithms used in our model.

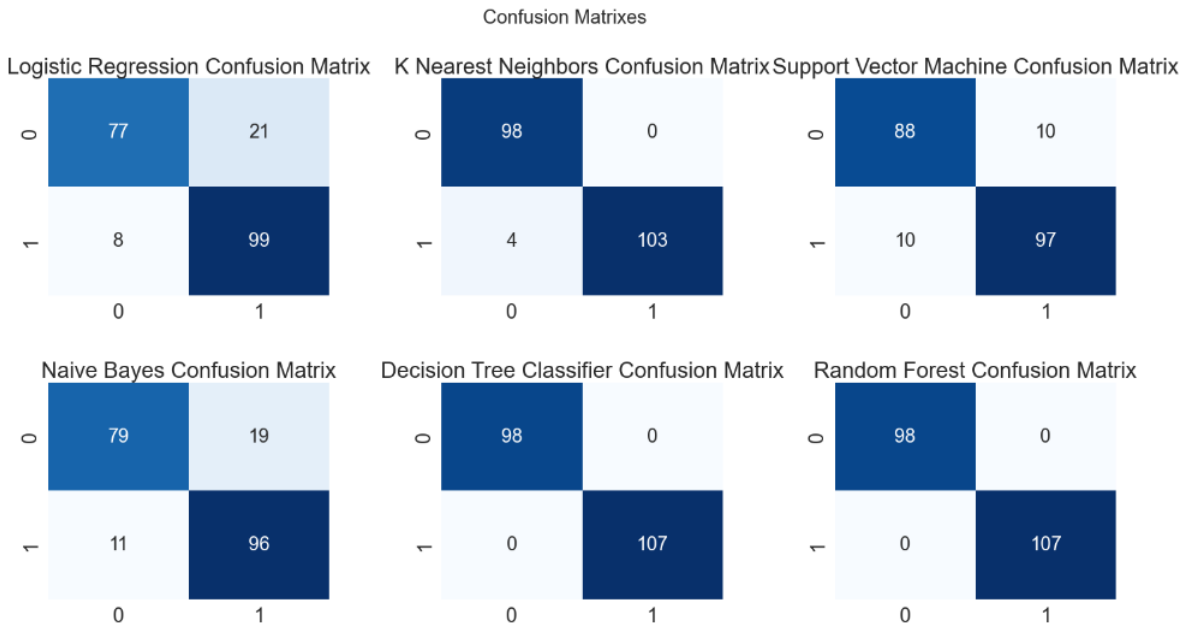


Figure 8.2: Confusion Matrix of Different Models

Chapter 9

Literature Review/ Related Works

Researchers in this area have developed techniques to predict cardiovascular disease using supervised machine learning algorithms. On this subject, several research papers have been prepared. Some of them related to our topic is given below;

Their study concentrated on using machine learning techniques to create an AI-based system for predicting diseases such as heart disease.They show how machine learning can help in cardiac illness prediction. They developed a python-based application for healthcare research since it is more reliable and helps in tracking and establishing various kinds of health monitoring apps. It is shown by their model how to convert classified columns in data processing and work with explanatory data. The important phases of application development are discussed, including the gathering of databases, the use of logistic regression, and the examination of the characteristics of the data set. In order to better diagnose cardiac issues, a random forest categorization system is being built.Data analysis is necessary for this application, which is regarded as important because to its about 83 percent correctness rate compared to training data. Then, the random forest classifier algorithm is explained together with the outcomes and experiments that increase the quality of research diagnosis. Goals, restrictions, and research contributions are included in this report's conclusion [2].

The purpose of this research is to use Python to look for cardiac issues. This project also utilized a number of additional import libraries, such as matplotlib, Numpy, Pandas, warnings, and numerous others. The results of the needed data set were evaluated using the correlation matrix, histogram, support vector classifier, K Neighbors classifier, Random Forest classifier, and Decision Tree classifier, all by using the Python computer language. Another open-source language that promotes the creation of fresh approaches for the healthcare industry and offers better patient results, leading to improved care delivery, is Python.Arrhythmia, often known as atherosclerosis (the hardening of the arteries brought on by an irregular heartbeat), is hence a type of cardiac disease. These signs and symptoms can occur in heart attack sufferers. Additionally, you can experience arm pain, light headedness, sore throat, snoring, and perspiration. Heart attacks, strokes, and coronary heart disease, sometimes referred to as heart failure and coronary artery disease, affect people over 65 much more frequently than they do people under 65 [2].

Many academics have previously proposed to employ machine learning algorithm while predicting and generating a set of output from desired data sets. The comparison approaches used are confusion matrix, precision, specificity, sensitivity, and F1 score. The K-Neighbors classifier outperformed the ML approach for the 13 attributes in the data set when data preprocessing was applied. The results are promising, and in the third technique, the data set was normalized while taking into account outliers and feature selection. The results are significantly superior to those obtained by previous methods [1].

Despite the large number of classifiers available, random forest classifiers offer the best accuracy for this project. Age, gender, blood pressure, cholesterol, and obesity are just a few of the medical traits used by this project to forecast results. In addition, the EHDPS forecasts patients' risk for heart disease. Significant information, health reasons, links associated to heart disease, and trends can be identified as a result of this project's work. The random forest method is used in this research to diagnose cardiac disease. This algorithm is applied in the Python approach for detecting heart disease. The accuracy rate of 83 percent (approximately) over training data is regarded significant for this application [2].

The research conducted an experiment in which different data mining algorithms were used to predict heart attacks and the best method of prediction was compared. When employing several categorization methods in data mining, the research results show no significant differences in prediction. The research could be a useful tool for doctors to predict dangerous instances in their practice and to provide correct guidance. The classification model will be able to answer more complicated questions about heart attack disease prediction [16].

This research examines a variety of machine learning methods, including K nearest neighbors (KNN), Logistic Regression, and Random Forest Classifiers, all of which can assist practitioners or medical professionals for effectively diagnosing Heart Disease. Using various classifiers, an effective Heart Disease Prediction System (EHDPS) has been built in this model. For prediction, this model incorporates 13 medical characteristics including chest pain, fasting sugar, blood pressure, cholesterol, age, and sex [12].

This research looks at the predictive accuracy of various machine learning approaches for estimating cardiovascular risk. The feasibility and utility of several machine learning techniques are investigated in this research. The proposed CDPS mission is to use machine learning techniques to aid professionals in making informed decisions and suggestions. The suggested CDPS was assessed using a variety of measures in order to find the optimal machine learning model. When it comes to predicting patients with cardiovascular disease, the Random Tree model performed extraordinarily well, with a 100% accuracy rate [20].

Machine learning techniques were used in this study to examine the raw data and provide a fresh and unique discernment for heart disease. The proposed hybrid HRFLM approach combines Random Forest (RF) and Linear Method features (LM) in this research. Using HRFLM, it was demonstrated that it was fairly reliable at predicting heart disease. [18] .

In this research, they used Machine Learning Algorithms on a cardiovascular disease data set to predict patients who have a long-term cardiovascular disease from those who are not, based on the information for each patient's feature. The goal was to consider numerous arrangement models and determine which was the most productive. The K-Nearest Neighbor, Support Vector Machines, Logistic Regression, Naive Bayes, and Random Forest algorithms were used in the research. The results of the re-enactment showed that the Logistic Regression classifier excelled at foreseeing with the best precision and shortest execution time [27] .

The Random Forest algorithm outperforms the Support Vector Machine, Logistic Regression, Naive Bayes, and Support Vector Machine algorithms in terms of accuracy. That is why they used random forest in this study. Based on the specifics of the diabetes-heart disease association, this algorithm can also be used to determine the likelihood of acquiring heart disease. The goal of the suggested model is to collect significant data on all components associated with coronary heart disease and the characteristics that influence it, train the data using the proposed AI computation, and predict how likely a patient will develop a coronary heart disease [24].

The research aims to identify patients who are more likely to develop heart disease based on a variety of medical factors. Researchers developed a technique to predict the likelihood that a patient will be diagnosed with heart disease based on their medical history. They used a variety of machine learning techniques, including KNN, Random Forest, and logistic regression, to predict and categorize patients with heart disease. A very beneficial technique was used to provide rules for how the model can be utilized to enhance heart attack prediction accuracy for any individuals. The strength of the proposed model was pretty adequate, as it could predict signs of having a heart problem in a particular person using KNN, Logistic Regression, and Random Forest, all of which exhibited good accuracy when compared to previously used classifiers like Naive Bayes and others. Therefore, a significant amount of stress has been reduced by using the given model to calculate the likelihood that the classifier will correctly and accurately diagnose heart sickness [12].

The results demonstrate that KNN, Random Forest Classifier, and Logistic regression perform better than the many algorithms most academics use to predict patients diagnosed with Heart disease, including SVC, Decision tree, and Logistic regression.In comparison to past studies, their algorithms are faster, more accurate, more cost-effective because they consume less resources. Additionally, the combined accuracy of KNN and Logistic Regression is 88.5 percent, exceeding or nearly matching the accuracy of earlier studies. This study demonstrates that Logistic Regression and KNN outperform Random Forest Classifier in the prediction of individuals with heart disease. This reveals that KNN and Logistic Regression are better at spotting cardiac issues. [12].

This work was inspired by a substantial amount of research concerning the use of machine learning algorithms for the diagnosis of cardiovascular heart disease. An overview of the literature is provided in this paper. A reliable prediction of cardiovascular disease has been produced utilizing a variety of algorithms, including Logistic Regression, KNN, Random Forest Classifier, and others. Results show that each algorithm has a unique ability to capture the specified objectives [7].

This study explains the heart illness hypothesis and provides a machine learning-based method for identifying heart disease desires using a data set of heart disorders. They used cross-approval, three-element choice computations, seven classifier execution assessment metrics, and seven well-known machine learning algorithms.They have created a technique that can unquestionably distinguish healthy persons from those who have coronary disease. They have discussed the total number of classifiers used in the study as well as the feature certainty calculations, pre-planning strategies, recommendation techniques, and classifier execution assessment estimations. They claim that networks built on machine learning and with emotional support will make it easier for doctors to locate cardiac patients. [10].

The medical industry generates a large amount of data that has never been adapted. Here, new techniques are presented that lower costs and improve accurate heart illness prediction. The numerous research procedures considered in this study for the categorization and prediction of heart disease using ML and deep learning (DL) techniques are quite accurate in demonstrating the efficacy of these methods. [23].

In this research the author used modified k-means and Naive Bayes to predict cardiovascular problems. Diagnosing heart disease is a challenging process that requires excellent skill. Heart illness is indicated by a value of "1" for the attribute "Disease," whereas a value of "0" indicates that it is not present. Modified k-means is consistent with the categorical and combinational data was encountered in this instance. Using the two optimal parameters, they obtain the two farthest clusters. This models accuracy was measured by naive bayes. When identifying a heart condition, this predictor is 93 percent accurate, but only 89 percent accurate when establishing that a patient doesn't have one. [19].

The authors of this study looked into how k-Nearest Neighbors (kNN) might be used to detect cardiac illness. This study shows that, in terms of accuracy, kNN outperforms neural network ensemble. Contrary to decision trees, implementing integrated voting did not improve the accuracy of the kNN in the detection of patients with heart disease voting-based classifiers that improve accuracy. One of the method of accumulation is voting which combines the outcomes of several classifiers. The maximum accuracy was achieved by kNN without voting, at 97.4 percent. With voting, the accuracy for kNN decreased to 92.7 percent [26].

The researchers here suggested a data-based prototype for predicting cardiac disease. Which is a mining methods, particularly Naive Bayes. A statistical classifier called Naive Bayes assigns no dependence between the attributes. To identify the class, the predictive performance must be maximized. In this case, the Naive Bayes classifier also works well. Here in this research, naive Bayes appears to be the most effective model for disease prediction, followed by neural networks and decision trees. [14].

A Heart Disease Prediction System prototype was created by Shadab Adam Pattekari et al. using Naive Bayes, Decision Trees, and Neural Networks. It is put into action in an internet software. User input is required for this system's predetermined questions. In order to compare the user's values with the training data set, it first obtains hidden data from the stored database. From a historical heart illness database, the technology unearths and extracts hidden facts about heart diseases. It can respond to difficult questions regarding disease diagnosis. The Naive Bayes classification method was used to estimate the likelihood of having heart disease after choosing a set of 15 features. [21].

In this study, they divided clinical knowledge into five categories: nil, low, normal, excessive, and very excessive. The method will assign it to the appropriate class label if any unidentified samples are found. The collected data here is the coronary heart disease foundation data set from the Cleveland Medical Center, which consists of 303 observations and 14 parameters. Coaching phase and testing phase are the two phases in which the system operates. The classification is overseen in the coaching section, and the checking out segment involves the prediction of unknown information or deficient values. The Naive Bayes algorithm, which is deployed here, is founded on the Bayes theorem. The outcomes show that consistency was accomplished by adjusting the number of occurrences in the sample set of information. [17].

Chapter 10

Limitations

The process of making a medical diagnosis is regarded as being important but complicated, and it demands accuracy and efficiency. Mechanisation of the same process would be a great benefit.

In order to compare the best way of prediction, the research experimented by applying several data mining algorithms to the prediction of heart attacks. The research findings do not show a significant difference in the prediction when utilizing various categorization methods in data mining. Although ML-based algorithms appear to perform well, they are far from perfect. A number of methodological limitations can affect results and increase heterogeneity. Such as;

First, technical factors in algorithms, such as hyper parameter tuning, are rarely made public, which resulting in considerable statistical variability.

Second, the data partitioning is arbitrary due to the lack of standardized implementation rules.

Third, feature selection strategies and methodologies are arbitrary and diverse.

Fourth, we were unable to classify the type of custom-built algorithms due to their ambiguity.

Most importantly, certain analyses did not correspond to the clinical setting, which is making the interpretation more difficult [15] . Despite the fact that there number of hurdles to clear before ML algorithms may be used in clinical practice, overall, ML algorithms have shown encouraging outcomes.

Chapter 11

Early Stages of development

Before preparing this research paper we have studied various data sets and observed different types of modalities that are used to develop different type of models. A variety of algorithms are used as classifiers in different models. We have worked to generate the best and most accurate set of outputs for this model. A variety of illnesses that affect your heart are referred to as cardiovascular disease.

Unhealthy food, physical inactivity, nicotine use, and excessive alcohol consumption are the most major risk factors for heart disease and stroke [30]. Various harmful habits, such as excessive cholesterol, obesity, increased cholesterol level, hypertension, and so on, raise the risk of heart disease. However, as time passes, more research data and hospital patient records become available.

There are multiple open sources for gaining access to patient records, and studies can be undertaken to see how various computer technologies can be utilized to correctly diagnose people and detect this condition before it becomes fatal [1].

The study can be a useful tool for doctors to spot rising cases in their practice and offer suitable advice. The classification model will be able to respond to more complicated questions regarding the disorders that cause heart attacks.

Chapter 12

Future Goals

The long-term preservation of human life and the early identification of heart problems will benefit from the identification of the processing of raw healthcare data and cardiovascular information. Risk prediction methodologies are definitely beyond their prime, and more effort and resources are required to collect more observational data with long follow-ups in order to develop population-specific models that address all of the concerns raised by current risk prediction models.

We are trying to build a model which will help us in the upcoming days. We are hoping that our prediction model will result as more advanced population-specific modes based on more data and approaches which may become personalized risk assessments [6].

The association between heart disorders and other diseases can be calculated in order to create similar prediction systems. Additionally, new algorithms may be applied to boost accuracy. With more parameters employed in these algorithms, better performance is attained.

The model proposed in the research can be further extended and diversified by include more attributes in the data set. You can increase the amount of attributes used for prediction with better accuracy. Additionally, the data set's size can be expanded, this will also aid in obtaining preferred accuracy.

Chapter 13

Conclusion

Heart disease prediction, which employs a machine learning algorithm, gives consumers with a prediction result if the user has heart disease. Algorithms for machine learning have evolved as a result of recent technological advances. Since it has emerged as one of the leading causes of mortality, cardiovascular disease performance is an important issue in medical data analysis.

Machine learning has the potential to increase doctors' understanding, particularly in the foretelling of cardiac disease, enabling them to better accommodate patient diagnosis and therapy. In the research, different machine learning algorithms are examined for their effectiveness and usefulness.Identifying the risk of cardiovascular disease can help medical facilities determine which procedures or equipment are required. Our research is aimed to develop a prediction model that will help us to identify and understand these diseases better.

The forecasting and classification model is used in this study to analyze the conditions of patients with cardiovascular disorders. The classification models were then utilized based on the predicted results, dividing each vital sign into low, normal, and high categories. The research reveals a model that analyzes data to forecast vital signs and classify them into unique labels, which allowing health workers to make quick decisions. The research found that implementing the forecasting model to related activities results in high accuracy and enhanced performance.

In this research, six ML classification modeling techniques have been used to construct a model for the detection of cardiovascular disease. By extracting the patient medical history that results in a fatal heart illness from a data set that contains patients' medical histories such as chest pain, blood sugar levels, blood pressure, etc. More effective outcomes can be obtained by increasing the data set size, deep learning, and a variety of additional optimizations.

To further improve the evaluation findings, machine learning and several other optimization approaches can also be applied. The data can be normalized in more ways, and the results can be compared. Furthermore, there may be further approaches to combine Cardiovascular Disease-trained ML models with specific multimedia for the benefit of patients and medical professionals.

Bibliography

[1] Rohit Bharti, Aditya Khamparia, Mohammad Shabaz, Gaurav Dhiman, Sagar Pande, and Parneet Singh. Prediction of heart disease using a combination of machine learning and deep learning. *Computational intelligence and neuroscience*, 2021, 2021.

[2] Victor Chang, Vallabhanent Rupa Bhavani, Ariel Qianwen Xu, and MA Hossain. An artificial intelligence model for heart disease detection using machine learning algorithms. *Healthcare Analytics*, 2:100016, 2022.

[3] Deep Checks. Test set in machine learning. https://deepchecks.com/glossary/test-set-in-machine-learning/, 2022. [Online, accessed on September-2022].

[4] Sruthi ER. Understanding random forest. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/, 2022. [Online, accessed on September-2022].

[5] Cloud Factory. The essential guide to quality training data for machine learning. https://www.cloudfactory.com/training-data-guide, 2022. [Online, accessed on September-2022].

[6] Farshad Farzadfar. Cardiovascular disease risk prediction models: challenges and perspectives. *The LANCET Global Health*, 7(10):e1288–e1289, 2019.

[7] Andrea Ganna, Patrik KE Magnusson, Nancy L Pedersen, Ulf de Faire, Marie Reilly, Johan Ärnlöv, Johan Sundström, Anders Hamsten, and Erik Ingelsson. Multilocus genetic risk scores for coronary heart disease prediction. *Arteriosclerosis, thrombosis, and vascular biology*, 33(9):2267–2272, 2013.

[8] Mursal Furqan GL. *DETECTING AND RELIEVING ANXIETY ON-THE-GO USING MACHINE LEARNING TECHNIQUES*. PhD thesis, MEHRAN UNIVERSITY, 2020.

[9] Rati Goel. Heart disease prediction using various algorithms of machine learning. *Available at SSRN 3884968*, 2021.

[10] Amin Ul Haq, Jian Ping Li, Muhammad Hammad Memon, Shah Nazir, and Ruinan Sun. A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. *Mobile Information Systems*, 2018, 2018.

[11] Yunxing Jiang, Xianghui Zhang, Rulin Ma, Xinping Wang, Jiaming Liu, Mulatibieke Keerman, Yizhong Yan, Jiaolong Ma, Yanpeng Song, Jingyu Zhang,

et al. Cardiovascular disease prediction by machine learning algorithms based on cytokines in kazakhs of china. *Clinical epidemiology*, 13:417, 2021.

[12] Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain, and Preeti Nagrath. Heart disease prediction using machine learning algorithms. In *IOP conference series: materials science and engineering*, volume 1022, page 012072. IOP Publishing, 2021.

[13] Amal Joby. What is training data? how it's used in machine learning. https://learn.g2.com/training-data, 2021. [Online, accessed on September-2022].

[14] Kamal Kant and Kanwal Garg. Review of heart disease prediction using data mining classifications. *International Journal for Scientific Research & Development (IJSRD)*, 2(04):2321–0613, 2014.

[15] Chayakrit Krittanawong, Hafeez Ul Hassan Virk, Sripal Bangalore, Zhen Wang, Kipp W Johnson, Rachel Pinotti, HongJu Zhang, Scott Kaplin, Bharat Narasimhan, Takeshi Kitai, et al. Machine learning prediction in cardiovascular diseases: a meta-analysis. *Scientific Reports*, 10(1):1–11, 2020.

[16] Hlaudi Daniel Masethe and Mosima Anna Masethe. Prediction of heart disease using classification algorithms. In *Proceedings of the world Congress on Engineering and computer Science*, volume 2, pages 25–29, 2014.

[17] Dhanashree S Medhekar, Mayur P Bote, and Shruti D Deshmukh. Heart disease prediction system using naive bayes. *Int. J. Enhanced Res. Sci. Technol. Eng*, 2(3), 2013.

[18] Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivastava. Effective heart disease prediction using hybrid machine learning techniques. *IEEE access*, 7:81542–81554, 2019.

[19] Sairabi H Mujawar and PR Devale. Prediction of heart disease using modified k-means and by using naive bayes. *International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization)*, 3(10):10265–10273, 2015.

[20] Rajkumar Gangappa Nadakinamani, A Reyana, Sandeep Kautish, AS Vibith, Yogita Gupta, Sayed F Abdelwahab, and Ali Wagdy Mohamed. Clinical data analysis for prediction of cardiovascular disease using machine learning techniques. *Computational Intelligence and Neuroscience*, 2022, 2022.

[21] Shadab Adam Pattekari and Asma Parveen. Prediction system for heart disease using naïve bayes. *International Journal of Advanced Computer and Mathematical Sciences*, 3(3):290–294, 2012.

[22] Java Point. Train and test datasets in machine learning. https://www.javatpoint.com/train-and-test-datasets-in-machine-learning, 2022. [Online, accessed on September-2022].

[23] DK Ravish, KJ Shanthi, Nayana R Shenoy, and S Nisargh. Heart function monitoring, prediction and prevention of heart attacks: Using artificial neural

networks. In *2014 International Conference on Contemporary Computing and Informatics (IC3I)*, pages 1–6. IEEE, 2014.

[24] PE Rubini, CA Subasini, A Vanitha Katharine, V Kumaresan, S Gowdham Kumar, and TM Nithya. A cardiovascular disease prediction using machine learning algorithms. *Annals of the Romanian Society for Cell Biology*, pages 904–912, 2021.

[25] Anshul Saini. Decision tree algorithm – a complete guide. https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/, 2021. [Online, accessed on September-2022].

[26] Mai Shouman, Tim Turner, and Rob Stocker. Applying k-nearest neighbour in diagnosing heart disease patients. *International Journal of Information and Education Technology*, 2(3):220–223, 2012.

[27] Muktevi Srivenkatesh. Prediction of cardiovascular disease using machine learning algorithms. *Int. J. Eng. Adv. Technol*, 9(3):2404–2414, 2020.

[28] Techopedia. Test set. https://www.techopedia.com/definition/33279/test-set, 2022. [Online, accessed on September-2022].

[29] Techopedia. Training data. https://www.techopedia.com/definition/33181/training-data, 2022. [Online, accessed on September-2022].

[30] WHO. Cardiovascular diseases. https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1, 2022. [Online, accessed on September-2022].

[31] Wikipedia. Training, validation, and test data sets. https://en.wikipedia.org/wiki/Training,_validation,_and_test_data_sets, 2022. [Online, accessed on September-2022].