# Heart Disease Prediction using Machine Learning Algorithms


## By

**Bishal Ghosh**
**Roll no: 25500115014**
**Registration no: 152550110015**
**Chayan Datta**
**Roll no: 25500115015**
**Registration no: 152550110016**
**Nilotpal Majumdar**
**Roll no: 25500115024**
**Registration no: 1525500110026**
**Sayan Paul**
**Roll no: 25500115038**
**Registration no: 152550110040**

## Under the Guidance of
## Mrs. Debdatta Chatterjee

**Bachelor of Technology(Computer Science & Engineering)**
**Department of Computer Science & Engineering**
**Dr. Sudhir Chandra Sur Degree Engineering College**
**Maulana Abul Kalam Azad Institute of Technology**
**Kolkata, West Bengal**

# CERTIFICATE

This is to certify that this is a bona fide record to the FINAL YEAR project work "Heart Disease Prediction using Machine Learning" done satisfactorily at DR. SUDHIR CHANDRA SUR DEGREE ENGINEERING COLLEGE by BISHAL GHOSH, CHAYAN DATTA, NILOTPAL MAJUMDAR, SAYAN PAUL of 8th SEMESTER, CSE.

This report or similar report on this topic has not been submitted for any other examination and does not form part of any other course undergone by the candidate. I have no doubt that they have a very good research potential.

I would like to wish them a very bright future.

**Date: _____**

_____

**Mrs. Debdatta Chatterjee**
**(**Prof, CSE, DSCSDEC)

_____

**Mrs. Mallika De**
(HOD,CSE,DSCSDEC)

_____

**(**EXAMINER**)**

# ACKNOWLEDGEMENT

We would like to take this opportunity to express our gratitude towards all those people who have in various ways, helped in the successful completion of our final year project work on "**Heart Disease Prediction using Machine Learning Algorithms**" done satisfactorily at DR. SUDHIR CHANDRA SUR DEGREE ENGINEERING COLLEGE. We must convey our gratitude to **Mrs. Debdatta Chatterjee** and **Mr. Sovan Saha** for giving us the constant source of inspiration and help in preparing the project, personally correcting our work and providing encouragement throughout the project. In that regard we are especially thankful to our Head of the Department **Mrs. Mallika De** for steering me through the tough as well as easy phases of the project in a result-oriented manner with concern attention. A special thanks to all the faculty members of our CSE department.

We would also love to take this opportunity to tell the readers of our material that their comments, be it appreciation or criticism would be the most valuable thing to us. That will be something, we will always be thankful for.

**Date: _____**

# ABSTRACT

Nowadays, health disease are increasing day by day due to lifestyle, hereditary. Especially, heart disease has become more common these days, i.e. life of people is at risk. Each individual has different values for Blood pressure, cholesterol and pulse rate. But according to medically proven results the normal values of Blood pressure is 120/90, Cholesterol is 100-129 mg/dL ,Pulse rate is 72, Fasting Blood Sugar level is 100 mg/dL ,Heart rate is 60-100 bpm, ECG is normal, Width of major vessels is 25 mm (1 inch) in the aorta to only 8 μm in the capillaries. This paper gives the survey about different classification techniques used for predicting the risk level of each person based on age, gender, Blood pressure, cholesterol, pulse rate.

"Disease Prediction" system based on predictive modeling predicts the disease of the user on the basis of the symptoms that user provides as an input to the system. The system analyzes the symptoms provided by the user as input and gives the probability of the disease as an output. Disease Prediction is done by implementing 5 techniques such as Naïve Bayes, KNN, Decision Tree, Linear Regression and Random Forest Algorithms. These techniques calculate the probability of the disease. Therefore, average prediction accuracy probability 83% is obtained.

# CONTENTS

## CHAPTER 1: INTRODUCTION

# CHAPTER: 1 INTRODUCTION

## 1.1 Introduction

In day to day life many factors that affect a human heart. Many problems are occurring at a rapid pace and new heart diseases are rapidly being identified. In today's world of stress Heart, being an essential organ in a human body which pumps blood through the body for the blood circulation is essential and its health is to be conserved for a healthy living. The health of a human heart is based on the experiences in a person's life and is completely dependent on professional and personal behaviors of a person. There may also be several genetic factors through which a type of heart disease is passed down from generations. According to the World Health Organization, every year more than 12 million deaths are occurring worldwide due to the various types of heart diseases which is also known by the term cardiovascular disease. The term Heart disease includes many diseases that are diverse and specifically affect the heart and the arteries of a human being. Even young aged people around their 20-30 years of lifespan are getting affected by heart diseases. The increase in the possibility of heart disease among young may be due to the bad eating habits, lack of sleep, restless nature, depression and numerous other factors such as obesity, poor diet, family history, high blood pressure, high blood cholesterol, idle behavior, family history, smoking and hypertension.

The diagnosis of the heart diseases is a very important and is itself the most complicated task in the medical field. All the mentioned factors are taken into consideration when analyzing and understanding the patients by the doctor through manual check-ups at regular intervals of time.

The symptoms of heart disease greatly depend upon which of the discomfort felt by an individual. Some symptoms are not usually identified by the common people. However, common symptoms include chest pain, breathlessness, and heart palpitations. The chest pain common to many types of heart disease is known as angina, or angina pectoris, and occurs when a part of the heart does not receive enough oxygen. Angina may be triggered by stressful events or physical exertion and normally lasts under 10 minutes. Heart attacks can also occur

as a result of different types of heart disease. The signs of a heart attack are like angina except that they can occur during rest and tend to be more severe. The symptoms of a heart attack can sometimes resemble indigestion. Heartburn and a stomach ache can occur, as well as a heavy feeling in the chest. Other symptoms of a heart attack include pain that travels through the body, for example from the chest to the arms, neck, back, abdomen, or jaw, lightheadedness and dizzy sensations, profuse sweating, nausea and vomiting.

Heart failure is also an outcome of heart disease, and breathlessness can occur when the heart becomes too weak to circulate blood. Some heart conditions occur with no symptoms at all, especially in older adults and individuals with diabetes. The term 'congenital heart disease' covers a range of conditions, but the general symptoms include sweating, high levels of fatigue, fast heartbeat and breathing, breathlessness, chest pain. However, these symptoms might not develop until a person is older than 13 years. In these types of cases, the diagnosis becomes an intricate task requiring great experience and high skill. A risk of a heart attack or the possibility of the heart disease if identified early, can help the patients take precautions and take regulatory measures. Recently, the healthcare industry has been generating huge amounts of data about patients and their disease diagnosis reports are being especially taken for the prediction of heart attacks worldwide. When the data about heart disease is huge, the machine learning techniques can be implemented for the analysis.

Data Mining is a task of extracting the vital decision-making information from a collective of past records for future analysis or prediction. The information may be hidden and is not identifiable without the use of data mining. The classification is one data mining technique through which the future outcome or predictions can be made based on the historical data that is available. The medical data mining made a possible solution to integrate the classification techniques and provide computerized training on the dataset that further leads to exploring the hidden patterns in the medical data sets which is used for the prediction of the patient's future state. Thus, by using medical data mining it is possible to provide insights on a patient's history and is able to provide clinical support through the analysis. For clinical analysis of the patients, these patterns are very much essential. In simple English, the medical data mining uses classification algorithms that is a vital part for identifying the possibility of heart attack before the occurrence. The classification algorithms can be trained and tested to make the predictions that determine the person's nature of being affected by heart disease[1].

In this research work, the supervised machine learning concept is utilized for making the predictions. A comparative analysis of the three data mining classification algorithms namely Random Forest, Decision Tree and Naïve Bayes are used to make predictions. The analysis is done at several levels of cross validation and several percentage of percentage split evaluation methods respectively. The Stat Log dataset from UCI machine learning repository is utilized for making heart disease predictions in this research work. The predictions are made using the classification model that is built from the classification algorithms when the heart disease dataset is used for training. This final model can be used for prediction of any types of heart diseases.[2][3]

Health-care is a field of the most needed service and an economically 2ndlargest industry in 21stcentury. While we talk about the affordability and quality assurance in health-care industry, several statistical analyses are carried on making health solutions more precise and flawless in this current era of increasing health problems and chronic diseases. Advancements on data driven intelligent technologies is disease diagnosis and detection, treatment and research are remarkable. Medical image analysis, symptom-based disease prediction is the part where the most sought-after brains are working. In this paper we aim to present our proposed model on the prediction on diagnosis of cardio vascular disease with ECG analysis and symptom-based detection. The model aims to be researched and advance in further to become robust and end to end reliable research tool. We will discuss about the classical methods and algorithms implemented on CVD prediction, gradual advancements, draw comparison of performance among the existing systems and propose an enhanced multi-module system performing better in terms of accuracy and feasibility. Implementation, training and testing of the modules have been done on datasets obtained from UCI and Physio net data repositories. Data format have been modified in case of the ECG report data for betterment of action by the convolutional neural network used in our research and in the risk prediction module we have chosen attributes for training and implementing the multi-layered neural network developed by us. The further research and advancement possibilities are also mentioned in the paper. [4]

## 1.2 Motivation

The main motivation of doing this research is to present a heart disease prediction model for the prediction of occurrence of heart disease. Further, this research work is aimed towards identifying the best classification algorithm for identifying the possibility of heart disease in a patient. This work is justified by performing a comparative study and analysis using three classification algorithms namely Naïve Bayes, Decision Tree, KNN, Logistic Regression and Random Forest are used at different levels of evaluations. Although these are commonly used machine learning algorithms, the heart disease prediction is a vital task involving highest possible accuracy. Hence, the three algorithms are evaluated at numerous levels and types of evaluation strategies. This will provide researchers and medical practitioners to establish a better understanding and help them identify a solution to identify the best method for predicting the heart diseases[5][6].

A key challenge confronting healthcare organization (hospitals, medical centers) is the facility of quality services at reasonable prices. Quality amenities suggest diagnosing patients accurately and regulating medications that are effective. Poor clinical choices can prompt deplorable results, which are in this manner unsatisfactory. Hospitals should limit the cost of clinical tests. They can accomplish these outcomes by utilizing fitting PC based data and additionally choice emotionally supportive networks [7][8]. The heart is the essential piece of our body. Life is itself reliant on effective working of the heart. If task of the heart isn't legitimate, it will influence the other body parts of human, for example, cerebrum, kidney and so on. Coronary illness is a sickness that effects on the activity of the heart. There are several elements which builds danger of Heart ailment [9].

Some of them are listed below:

• The family history of heart disease

• Smoking

• Cholesterol

• High blood pressure

• Obesity

• Lack of physical exercise

Because of a wide accessibility of superlative measure of information and a need to change over this accessible huge measure of information to helpful data requires the utilization of information mining strategies. Information Mining and KDD (learning disclosure in the database) have turned out to be prominent as of late[10][11]. The popularity of information mining and KDD (information revelation in database) shouldn't be an amazement since the measure of the information increases that are accessible are extremely extensive to be analyzed physically and even the techniques for programmed information investigation in view of established insights and machine adapting frequently threaten issues when preparing large, dynamic information increases comprising of complex items [12]. Information Mining is the center piece of Knowledge Discovery Database (KDD). Numerous individuals regard Data Mining as an equivalent word for KDD since it's a key piece of KDD process[13][14]. There are sure stages of information mining that you will need to get comfortable with, and these are exploration, pattern identification, and deployment. Information mining is an iterative procedure that commonly includes the accompanying stage [15].

## 1.3 Literature Review

According to Ordonez [16]the heart disease can be predicted with some basic attributes taken from the patient and in their work have introduced a system that includes the characteristics of an individual human being based on totally 13 basic attributes like sex, blood pressure, cholesterol and others to predict the likelihood of a patient getting affected by heart disease. They have added two more attributes i.e. fat and smoking behavior and extended the research dataset. The data mining classification algorithms such as Decision Tree, Naive Bayes, and Neural Network are utilized to make predictions and the results are analyzed on Heart disease database.

Yilmaz, [17]have proposed a method that uses least squares support vector machine (LS-SVM) utilizing a binary decision tree for classification of cardiotocogram to find out the patient condition.

Duff, et al. [18]have done a research work involving five hundred and thirty-three patients who had suffered from cardiac arrest and they were integrated in the analysis of heart disease probabilities. They performed classical statistical analysis and data mining analysis using mostly Bayesian networks.

Frawley, et al. [19] have performed a work on prediction of survival of Coronary heart disease (CHD) which is a challenging research problem for medical society. They also used 10-fold cross-validation methods to determine the impartial estimate of the three prediction models for performance comparison purposes.

Lee et al.[20] proposed a novel methodology to expand and study the multi-parametric feature along with linear and nonlinear features of Heart Rate Variability diagnosing cardiovascular disease. They have carried out various experiments on linear and non-linear features to estimate several classifiers, e.g., Bayesian classifiers, CMAR, C4.5 and SVM. Based on their experiments, SVM outperformed the other classifiers.

Noh et al. [21]suggested a classification method which is an associative classifier that is constructed based on the efficient FP-growth method. Because the volume of patterns can be diverse and huge, they offered a rule to measure the cohesion and in turn allow a tough choice of pruning patterns in the pattern-generating process.

Parthiban, et al. [22] have proposed a new work in which the heart disease is identified and predicted using the proposed Coactive Neuro-Fuzzy Inference System (CANFIS). Their model works based on the collective nature of neural network adaptive capabilities and based on the genetic algorithm along with fuzzy logic in order to diagnose the occurrence of the disease. The performance of the proposed CANFIS model was evaluated in terms of training performances and classification accuracies. Finally, their results show that the proposed CANFIS model has great prospective in predicting the heart disease.

Singh, et al. [23] have done a work using, one partition clustering algorithm (K-Means) and one hierarchical clustering algorithm (agglomerative). K-means algorithm has higher effectiveness and scalability and converges fast when production with large data sets. Hierarchical clustering constructs a hierarchy of clusters by either frequently merging two smaller clusters into a larger one or splitting a larger cluster into smaller ones. Using WEKA data mining tool, they have calculated the performance of k-means and hierarchical clustering algorithm on the basis of accuracy and running time.

Guru, et al. [24] have proposed the computational model based on a multilayer perceptron with three layers is employed to enlarge a decision support system for the finding of five major heart diseases. The proposed decision support system is trained using a back propagation algorithm amplified with the momentum term, the adaptive learning rate and the forgetting mechanics.

Palaniappan, et al. [25] have carried out a research work and have built a model known as Intelligent Heart Disease Prediction System (IHDPS) by using several data mining techniques such as Decision Trees, Naïve Bayes and Neural Network.

Shantakumar, et al. [26] have done a research work in which the intelligent and effective heart attack prediction system is developed using Multi-Layer Perceptron with Back-Propagation. Accordingly, the frequency patterns of the heart disease are mined with the MAFIA algorithm based on the data extracted.

Yanwei, et.al [27] have built a classification method based on the origin of multi parametric features by assessing HRV (Heart Rate Variability) from ECG and the data is pre-processed and heart disease prediction model is built that classifies the heart disease of a patient.

Data mining plays an important role in the field of heart disease prediction. [28] Medical Data mining has great potential like exploring the hidden patterns which can be utilized for clinical diagnosis of any disease dataset[29]. Several data mining techniques are used in the diagnosis of heart disease such as Naive Bayes, Decision Tree, neural network, kernel density, bagging algorithm, and support vector machine showing different levels of accuracies. Naive Bayes is one of the successful classification techniques used in the diagnosis of heart disease patients. Peter et al. [30] talked about a new feature selection method algorithm which is the hybrid method which combined CFS and Bayes theorem (CFS+Filter Subset Eval) and evaluated accuracy 85.5%.

Shouman [31] presented work by integrating k-means clustering with Naive Bayes using different initial centroid selection to improve the Naive Bayes accuracy for diagnosing heart disease patients and accuracy was 84.5%. Rupali et al. [32] decision support in Heart Disease Prediction

System (HDPS) is developed by using both Naive Bayesian Classification and Jelinek-Mercer smoothing technique. This Laplace smoothing is used to make an approximating function which attempts to capture important patterns in the data to avoid noise & accuracy is 86%. Elma et al. [33] proposed a classifier with the distance-based algorithm K-nearest neighbor and

statistical based Naïve Bayes classifier (cNK) and achieved the accuracy 85.92% for heart disease dataset.[34]

Data mining has been played an important role in the intelligent medical systems [35][36]. The relationships of disorders and the real causes of the disorders and the effects of symptoms that are spontaneously seen in patients can be evaluated by the users via the constructed software easily. Large databases can be applied as the input data to the software by using the extendibility of the software. The effects of relationships that have not been evaluated adequately have been explored and the relationships of hidden knowledge laid among the large medical databases have been searched in this study by means of finding frequent items using candidate generation. The sets of sicknesses simultaneously seen in the medical databases can be reduced by using our non-candidate approach. Knowledge of the risk factors associated with heart disease helps health care professionals to identify patients at high risk of having heart disease. Statistical analysis and data mining techniques to help healthcare professionals in the diagnosis of heart disease. Statistical analysis has identified the disorders of the heart and blood vessels, and includes coronary heart disease (heart attacks), cerebrovascular disease (stroke), raised blood pressure (hypertension), peripheral artery disease, rheumatic heart disease, congenital heart disease and heart failure. The major causes of cardiovascular disease are tobacco use, physical inactivity, an unhealthy diet and harmful use of alcohol. The three major causes of heart diseases are chest pain, stroke and heart attack [37]. The data mining methods like artificial neural network technique is used in effective heart attack prediction system. First the dataset used for prediction of heart diseases was pre-processed and clustered by means of K-means clustering algorithm [38]. Then neural network is trained with the selected significant patterns. Multi-layer Perceptron Neural Network with Back-propagation is used for training. The results indicate that the algorithm used is capable of predicting the heart diseases more efficiently. The prediction of heart diseases significantly uses 15 attributes, with basic data mining technique like ANN, Clustering and Association Rules, soft computing approaches etc. The outcome shows that Decision Tree performance is more and few time Bayesian classification is having similar accuracy as of decision tree but other predictive methods like K-Nearest Neighbor, Neural Networks, Classification based on clustering will not perform well [39]. By using the Weighted Associative Classifier (WAC), a slight change has been made, instead of considering 5 class labels, only 2 class labels are used. One for “Heart Disease” and another one for “No Heart Disease”. The maximum accuracy (81.51%) has been achieved.

When genetic algorithm is applied, the accuracy of the Decision Tree and Bayesian Classification is improved by reducing the actual data size. The dataset of 909 patient records with heart diseases has been collected and 13 attributes has been used for consistency [37]. The patient records have been splatted equally as 455 records for training dataset and 454 records for testing dataset. After applying genetic algorithm, the attributes has been reduced to 6 and decision tree performs more efficiently with 99.2% accuracy when compared with other algorithms. In 2011, Hnin Wint Khaing presented an efficient approach for the prediction of heart attack risk levels from the heart disease database. Firstly, the heart disease database is clustered using the K-means clustering algorithm, which will extract the data relevant to heart attack from the database. This approach allows mastering the number of fragments through its k parameter. Subsequently the frequent patterns are mined from the extracted data, relevant to heart disease, using the MAFIA (Maximal Frequent Item set Algorithm) algorithm. The machine learning algorithm is trained with the selected significant patterns for the effective prediction of heart attack. They have employed the ID3 algorithm as the training algorithm to show level of heart attack with the decision tree. The results showed that the designed prediction system is capable of predicting the heart attack effectively [40]. Chourasia and Pal conducted study on the prediction of heart attack risk levels from the heart disease data base. The prediction of heart diseases significantly uses 11important attributes, with basic data mining technique like Naïve Bayes, J48 decision tree and Bagging approaches. The outcome shows that bagging techniques performance is more accurate than Bayesian classification andJ48.The results shows that the bagging prediction system is capable of predicting the heart attack effectively[41]. Researchers have been applying various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., to help health care professionals with improved accuracy in the diagnosis of heart disease. The heart disease database used from the University of California Irvine. UCI archive is used. This database contains four data sets from the Cleveland Clinic Foundation, Hungarian Institute of Cardiology, V.A. Medical Center and University Hospital of Switzerland. However, here we discuss the Cleveland Heart Disease Dataset (CHDD). Data Mining Techniques. This paper uses data mining algorithms CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and decision table (DT). These classification algorithms are selected because they are very often used for research purposes and have potential to yield good results.

Moreover, they use different approaches for generating the classification models, which increases the chances for finding a prediction model with high classification accuracy. CART.

## 1.4 What is Heart Disease?

Heart disease describes a range of conditions that affect your heart. Diseases under the heart disease umbrella include blood vessel diseases, such as coronary artery disease; heart rhythm problems (arrhythmias); and heart defects you're born with (congenital heart defects), among others. The term "heart disease" is often used interchangeably with the term "cardiovascular disease." Cardiovascular disease generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain (angina) or stroke. Other heart conditions, such as those that affect your heart's muscle, valves or rhythm, also are considered forms of heart disease. Heart failure is a serious condition with high prevalence (about2% in the adult population in developed countries, and or than8%inpatients olderthan75years). About 3 – 5%ofhospitaladmissions are linked with heart failure incidents. Heart failure is the first cause of admission by healthcare professionals in their clinical practice. The costs are very high, reaching up to 2% of the total health costs in the developed countries. Building an effective disease management strategy requires analysis of large amount of data, early detection of the disease, assessment of the severity and early prediction of adverse events. This will inhibit the progression of the disease, will improve the quality of life of the patients and will reduce the associated medical costs. Toward this direction machine learning techniques have been employed. The aim of this paper is to present the state-of-the-art of the machine learning methodologies applied for the assessment of heart failure. More specifically, models predicting the presence, estimating the subtype, assessing the severity of heart failure and predicting the presence of adverse events, such as destabilizations, re-hospitalizations, and mortality are presented. According to the authors' knowledge, it is the first time that such a comprehensive review, focusing on all aspects of the management of heart failure, is presented.

### 1.4.1 How the Heart Works?

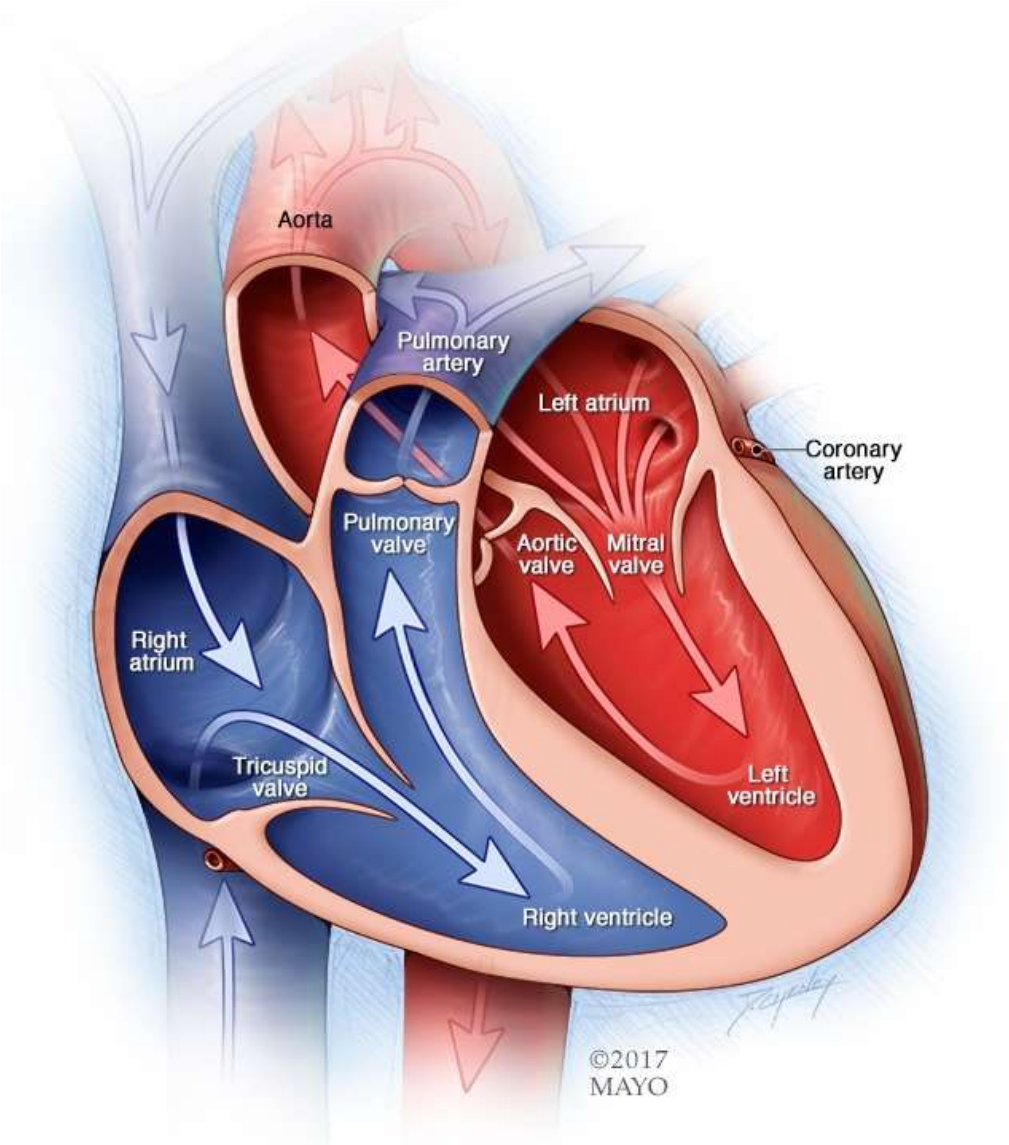



Figure 1 Working of the Heart

### 1.4.2 Chambers and valves of the heart

Your heart is a pump. It's a muscular organ about the size of your fist, situated slightly left of center in your chest. Your heart is divided into the right and the left side. The division prevents oxygen-rich blood from mixing with oxygen-poor blood. Oxygen-poor blood returns to the heart after circulating through your body.

- The right side of the heart, comprising the right atrium and ventricle, collects and pumps blood to the lungs through the pulmonary arteries.
- The lungs refresh the blood with a new supply of oxygen. The lungs also breathe out carbon dioxide, a waste product.
- Oxygen-rich blood then enters the left side of the heart, comprising the left atrium and ventricle.
- The left side of the heart pumps blood through the aorta to supply tissues throughout the body with oxygen and nutrients.

### 1.4.3 Heart valves

Four valves within your heart keep your blood moving the right way by opening only one way and only when they need to. To function properly, the valve must be formed properly, must open all the way and must close tightly so there's no leakage. The four valves are:

- Tricuspid
- Mitral
- Pulmonary
- Aortic

### 1.4.4 Causes of cardiovascular disease

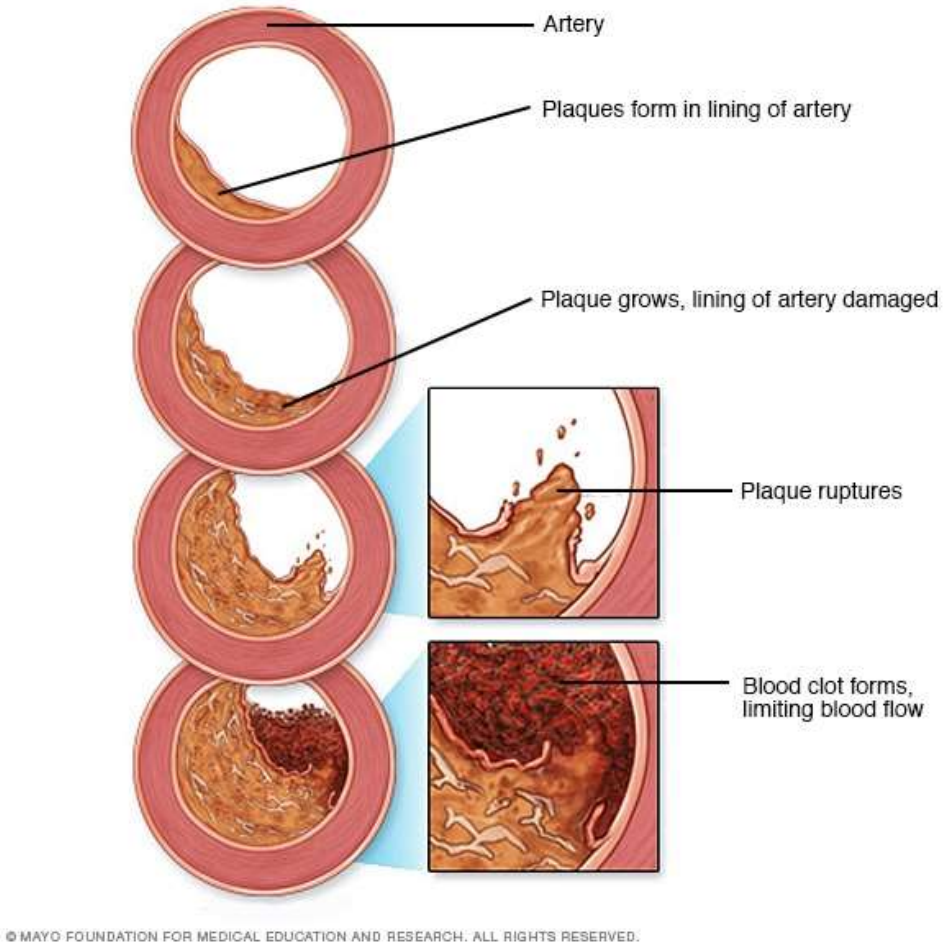



Figure 2 Causes of Cardiovascular Diseases

**Development of atherosclerosis -** While cardiovascular disease can refer to different heart or blood vessel problems, the term is often used to mean damage to your heart or blood vessels by atherosclerosis (ath-ur-o-skluh-ROE-sis), a buildup of fatty plaques in your arteries. Plaque buildup thickens and stiffens artery walls, which can inhibit blood flow through your arteries to your organs and tissues. Atherosclerosis is also the most common cause of cardiovascular disease. It can be caused by correctable problems, such as an unhealthy diet, lack of exercise, being overweight and smoking.

**Causes of heart arrhythmia -** Common causes of abnormal heart rhythms (arrhythmias) or conditions that can lead to arrhythmias include:

- Heart defects you're born with (congenital heart defects)
- Coronary artery disease
- High blood pressure
- Diabetes
- Smoking
- Excessive use of alcohol or caffeine
- Drug abuse
- Stress
- Some over-the-counter medications, prescription medications, dietary supplements and herbal remedies
- Valvular heart disease

In a healthy person with a normal, healthy heart, it's unlikely for a fatal arrhythmia to develop without some outside trigger, such as an electrical shock or the use of illegal drugs. That's primarily because a healthy person's heart is free from any abnormal conditions that cause an arrhythmia, such as an area of scarred tissue. However, in a heart that's diseased or deformed, the heart's electrical impulses may not properly start or travel through the heart, making arrhythmias more likely to develop.

**Causes of congenital heart defects -** Congenital heart defects usually develop while a baby is in the womb. Heart defects can develop as the heart develops, about a month after conception, changing the flow of blood in the heart. Some medical conditions, medications and genes may play a role in causing heart defects. Heart defects can also develop in adults. As you age, your heart's structure can change, causing a heart defect.

**Causes of cardiomyopathy -** The cause of cardiomyopathy, a thickening or enlarging of the heart muscle, may depend on the type:

- **Dilated cardiomyopathy.** The cause of this most common type of cardiomyopathy often is unknown. It may be caused by reduced blood flow to the heart (ischemic heart disease) resulting from damage after a heart attack, infections, toxins and certain drugs. It may also be inherited from a parent. It usually enlarges (dilates) the left ventricle.
- **Hypertrophic cardiomyopathy.** This type, in which the heart muscle becomes abnormally thick, usually is inherited. It can also develop over time because of high blood pressure or aging.
- **Restrictive cardiomyopathy.** This least common type of cardiomyopathy, which causes the heart muscle to become rigid and less elastic, can occur for no known reason. Or it may be caused by diseases, such as connective tissue disorders, excessive iron buildup in your body

(hemochromatosis), the buildup of abnormal proteins (amyloidosis) or by some cancer treatments.

**Causes of heart infection -** A heart infection, such as endocarditis, is caused when an irritant, such as a bacterium, virus or chemical, reaches your heart muscle. The most common causes of heart infection include:

- Bacteria
- Viruses
- Parasites

**Causes of valvular heart disease -** There are many causes of diseases of your heart valves. You may be born with valvular disease, or the valves may be damaged by conditions such as:

- Rheumatic fever
- Infections (infectious endocarditis)
- Connective tissue disorders

### 1.4.5 Risk factors

Risk factors for developing heart disease include

- **Age.** Aging increases your risk of damaged and narrowed arteries and weakened or thickened heart muscle.
- **Sex.** Men are generally at greater risk of heart disease. However, women's risk increases after menopause.
- **Family history.** A family history of heart disease increases your risk of coronary artery disease, especially if a parent developed it at an early age (before age 55 for a male relative, such as your brother or father, and 65 for a female relative, such as your mother or sister).
- **Smoking.** Nicotine constricts your blood vessels, and carbon monoxide can damage their inner lining, making them more susceptible to atherosclerosis. Heart attacks are more common in smokers than in nonsmokers.
- **Certain chemotherapy drugs and radiation therapy for cancer.** Some chemotherapy drugs and radiation therapies may increase the risk of cardiovascular disease.
- **Poor diet.** A diet that's high in fat, salt, sugar and cholesterol can contribute to the development of heart disease.

- **High blood pressure.** Uncontrolled high blood pressure can result in hardening and thickening of your arteries, narrowing the vessels through which blood flows.
- **High blood cholesterol levels.** High levels of cholesterol in your blood can increase the risk of formation of plaques and atherosclerosis.
- **Diabetes.** Diabetes increases your risk of heart disease. Both conditions share similar risk factors, such as obesity and high blood pressure.
- **Obesity.** Excess weight typically worsens other risk factors.
- **Physical inactivity.** Lack of exercise also is associated with many forms of heart disease and some of its other risk factors, as well.
- **Stress.** Unrelieved stress may damage your arteries and worsen other risk factors for heart disease.
- **Poor hygiene.** Not regularly washing your hands and not establishing other habits that can help prevent viral or bacterial infections can put you at risk of heart infections, especially if you already have an underlying heart condition. Poor dental health also may contribute to heart disease.

### 1.4.6 Complications

Complications of heart disease include:

- **Heart failure.** One of the most common complications of heart disease, heart failure occurs when your heart can't pump enough blood to meet your body's needs. Heart failure can result from many forms of heart disease, including heart defects, cardiovascular disease, valvular heart disease, heart infections or cardiomyopathy.
- **Heart attack.** A blood clot blocking the blood flow through a blood vessel that feeds the heart causes a heart attack, possibly damaging or destroying a part of the heart muscle. Atherosclerosis can cause a heart attack.
- **Stroke.** The risk factors that lead to cardiovascular disease also can lead to an ischemic stroke, which happens when the arteries to your brain are narrowed or blocked so that too little blood reaches your brain. A stroke is a medical emergency — brain tissue begins to die within just a few minutes of a stroke.
- **Aneurysm.** A serious complication that can occur anywhere in your body, an aneurysm is a bulge in the wall of your artery. If an aneurysm bursts, you may face life-threatening internal bleeding.

- **Peripheral artery disease.** Atherosclerosis also can lead to peripheral artery disease. When you develop peripheral artery disease, your extremities — usually your legs — don't receive enough blood flow. This causes symptoms, most notably leg pain when walking (claudication).
- **Sudden cardiac arrest.** Sudden cardiac arrest is the sudden, unexpected loss of heart function, breathing and consciousness, often caused by an arrhythmia. Sudden cardiac arrest is a medical emergency. If not treated immediately, it is fatal, resulting in sudden cardiac death.

### 1.4.7 Prevention

Certain types of heart disease, such as heart defects, can't be prevented. However, you can help prevent many other types of heart disease by making the same lifestyle changes that can improve your heart disease, such as:

- Quit smoking
- Control other health conditions, such as high blood pressure, high cholesterol and diabetes
- Exercise at least 30 minutes a day on most days of the week
- Eat a diet that's low in salt and saturated fat
- Maintain a healthy weight
- Reduce and manage stress
- Practice good hygiene

## 1.5 Some of the attributes we used for Heart Disease Prediction and their correlation to CVD ( Cardiovascular Diseases )

Below we have explained some of the key attributes we have taken in to consideration in our dataset for predicting whether the given data leads to conclude the presence of heart disease.

These key attributes are the very facts that has been used in determining a presence of heart disease. Thus, here we shall be getting into deeper in checking how these factors relate to or even cause Heart Diseases or CVD.

### 1.5.1 Age as a Cardiovascular Risk Factor

According to the most recent estimates from United States, cardiovascular disease (CVD) death rates have declined but the disease burden still remains substantially high.[42] The risk of developing CVD is largely (75–90%) explained by the presence or absence of traditional CVD risk factors.[43] Age is a well known traditional risk factor, which is generally considered to be non-modifiable for obvious reasons. In this review, we discuss the common use of an individual's age in prediction of CVD incidence using different risk scores, examine whether age as a risk factor can be modified or not, discuss the methods used to evaluate long- and short-term CVD risk, appropriate communication of an individual's risk based on their age group and CVD risk, and conclude by discussing the influence of age on cardiac and vascular risk factors.

**Assessment of CVD risk using Age as part of Risk Scores** - With aging, there is an incremental acquisition of several CVD risk factors in an individual's lifespan. When these risk factors are incorporated in a multivariable regression model, age still remains an independent risk factor. There are several risk prediction scores currently available to assess an individual's risk of CVD, and all of them include 'age' as a predictor. Older age, as assessed by these risk scores, is associated with greater risk of CVD. Although there are several risk scores available, the Framingham Risk Score (FRS)[44] is one of the most-widely adopted screening tools in United States and is recommended by National Heart Lung and Blood Institute to assess an individual's CVD risk. Other risk scores which are tested in Britain, Scotland, New Zealand or China have not been formally tested in the United States. In addition to the traditional risk factors (age, gender, smoking, total cholesterol, HDL-cholesterol and systolic blood pressure which are part of FRS), risk scores developed in Britain and Scotland also incorporate family history and social deprivation as risk factors, and these additional variables marginally improve prediction of CVD risk over the FRS when applied to the British and the Scottish populations, respectively. The Reynolds risk score also includes age as a component and is constructed using a database of middle-aged American women and requires the additional measurements of C-reactive protein and HbA1c (in diabetics).[45] Lastly, the risk prediction score reported in prior European studies[46] and currently adopted by the Joint European societies[47] is based on models which predict CVD death, and therefore underestimates the burden of CVD by not including the non-fatal events. Note that although CVD death rates have declined in some developed European countries (quite similar to the trend in the United States), the overall CVD burden still remains high.[48]

**Age is an Independent Risk Factor for Cardiovascular Disease -** As discussed above, even after adjusting for traditional risk factors in a multivariable CVD prediction model, age remains a fundamental predictor of CVD risk.[48] However, when age and other risk factors are used jointly to examine an individual's future risk of CVD, it has been postulated that the contribution of age in the multivariable models may be a reflection of the intensity and the duration of exposure to other traditional CVD risk factors.[49] If this observation were true, avoidance of these other risk factors should result in a reduction of CVD risk associated with age per se. To examine this hypothesis prior studies from Framingham Heart Study have shown that the absence of each of these traditional risk factors is associated with a reduction in the risk of CVD even at an older age.[49] When the absence of multiple risk factors is factored

into an individual's CVD risk assessment, the reduction in CVD risk is further augmented. Similarly, using the Framingham cohort, investigators have observed that lower midlife blood pressure and total cholesterol levels, absence of glucose intolerance, smoking abstinence, higher education and female gender all predicted increased survival up to 85 years of age.[50] Additionally at an older age, the contribution of age to CVD risk prediction declines, in part because there is less time left for an individual to acquire other modifiable CVD risk factors. Therefore, age at any given point influences the assessment of both the short- and long-term CVD risk of an individual. The absence of these CVD risk factors not only prevents the development of CVD but also decreases the risk of age-associated co-morbidities and mortality.[51] In another prior study, after excluding individuals with cancer, cardiovascular disease and diabetes before 50 years of age, investigators followed the Framingham cohort to evaluate who was likely to reach 75 years of age. They concluded that smoking fewer cigarettes per day, lower systolic blood pressure, and higher forced vital capacity were associated with longevity in both sexes.[52] Moreover, these observations relating to presence and absence of traditional risk factors have also been confirmed in a population-based study in the Japanese cohort from the Honolulu Heart Program,[53] and the large scale, multiethnic and international InterHeart Study.[54] The Inter Heart study investigators also tested this hypothesis in a case-control fashion among all age groups and observed similar results for prevention of myocardial infarction.[55] Therefore, it is now well established that life expectancy of an individual is dependent on modification of traditional risk factors and age-associated risk of CVD can be minimized by correcting or avoidance of these risk factors. Though, it is important to note that risk factor modification is equally important for both young and older individuals and will decrease their subsequent risk of CVD.[56]

**Relative risk versus Absolute risk Assessment -** Current CVD risk assessment using Framingham risk score comprises of the traditional risk factors i.e. cholesterol (total and HDL), blood pressure, history of smoking and age.[57] While assessing risk of CVD, it is important that both short-term (10-year CVD risk) and long-term (>10 year) risk for CVD are evaluated, and communicated appropriately to an individual.[58] At a younger age, an individual with several CVD risk factors (i.e. smoker, increased cholesterol and high blood pressure) will have a lower absolute short-term risk (compared to an older individual with similar CVD risk factors), and the absolute risk increases as the person gets older. However, the relative risk remains relatively invariant throughout a person's lifespan provided other risk factors (except

age) do not change, and it may actually decrease over time. Similarly, an older individual with several risk factors will have a higher short-term absolute risk (compared to a younger individual with a similar risk factor profile) even though the relative risk may remain constant through the lifespan, provided there is no change in risk factors.[59]

**Communicating CVD Risk to Young and Old -** Communicating either short- or long-term CVD risk to a patient can be challenging and might over or under-estimate the importance of risk factor reduction and therefore impact how a person would react by changing lifestyle for future risk reduction. For example, communicating an overestimated relative risk to a young individual might result in emotional or financial stress (may require them to take medications) whereas communicating an under-estimated absolute risk may result in a lower level of motivation on the part of an individual to work towards changing his/her lifestyle to reduce CVD risk.[21] Present guidelines from Adult Treatment Panel (ATP-III) for treatment of high blood cholesterol appropriately incorporates both relative and absolute risk assessment aspects (as discussed above) for an individual and provides flexibility for discussion by a treating physician in primary prevention settings.[60] Prior investigators have cautioned treating physicians to distance themselves from communicating the magnified relative risk of an individual (compared to lower absolute risk) in order to achieve professionally desirable goals.[61]

**Influence of Age on Other Individual Risk Factors -** It is intuitive that if age is an independent risk factor for developing CVD, the lifetime risk of CVD for an individual would continue to increase with age. However, the lifetime risk for CVD is lower at age 70 than at age 50 years, for an individual whose lifestyle risk factors remains unchanged.[62] Similarly, lifetime risk of coronary artery disease, stroke, hypertension and heart failure does not continue to increase with age. One explanation for this observation is that there is shorter time period left for older individuals to develop the disease and a greater hazard of death due to competing causes. Other reasons are that those who live longer have inherent bias of lower burden of cardiovascular risk factors which lowers their risk of developing an event, or a genetic makeup with resistance to develop cardiovascular disease.[63] Framingham cohort enrolled individuals at their midlife (30–62yrs) primarily but Inter Heart study included some young participants (<40yrs) and both showed similar results that reduction or absence of risk factors is additive

and improves mortality. Consequently, it is important to note that screening for risk factors and advice about modifications of risk factors should start at an early age.[64]

**Influence of Individual Risk Factors on Age-associated CVD Risk -** A sex-specific analysis from Framingham cohort suggests about 11.9% ( men) to 40.3% (women) of age-associated CVD risk may be attributable to the concomitant burden of other CVD risk factors.[65] These estimates are based on comparing unadjusted regression coefficients for age with those obtained after adjusting for other CVD risk factors in multivariable models (systolic blood pressure, diabetes, total to high-density lipoprotein cholesterol ratio, history of smoking and body mass index).[66]

### 1.5.2 Gender differences in coronary heart disease

Although CVD remains the leading killer of both women and men in the United States, there are substantial sex/gender differences in the prevalence and burden of different CVDs, as outlined. For both women and men, coronary heart disease (CHD) is the largest contributor to CVD morbidity and mortality. The absolute numbers of women living with and dying of CVD and stroke exceed those of men, as does the number of hospital discharges for heart failure and stroke.[67] In 2007, women accounted for 60.6% of US stroke deaths.[68] In contrast, more men are living with and dying of CHD than women and have more hospital discharges for CVD and CHD. As shown in Figure 3, the prevalence of CHD is higher in men within each age stratum until after 75 years of age, which may contribute to the perception that heart disease is a man's disease. Sex differences in CVD and CHD mortality largely reflect sex differences in US demographics. Because female sex is associated with a longer life expectancy than male sex, women constitute a larger proportion of the elderly population in which the prevalence of CVD is greatest. Alarming statistics among younger women 35 to 44 years of age show that CHD mortality rates have increased an average of 1.3% annually between 1997 and 2002, a statistically significant trend.[69]

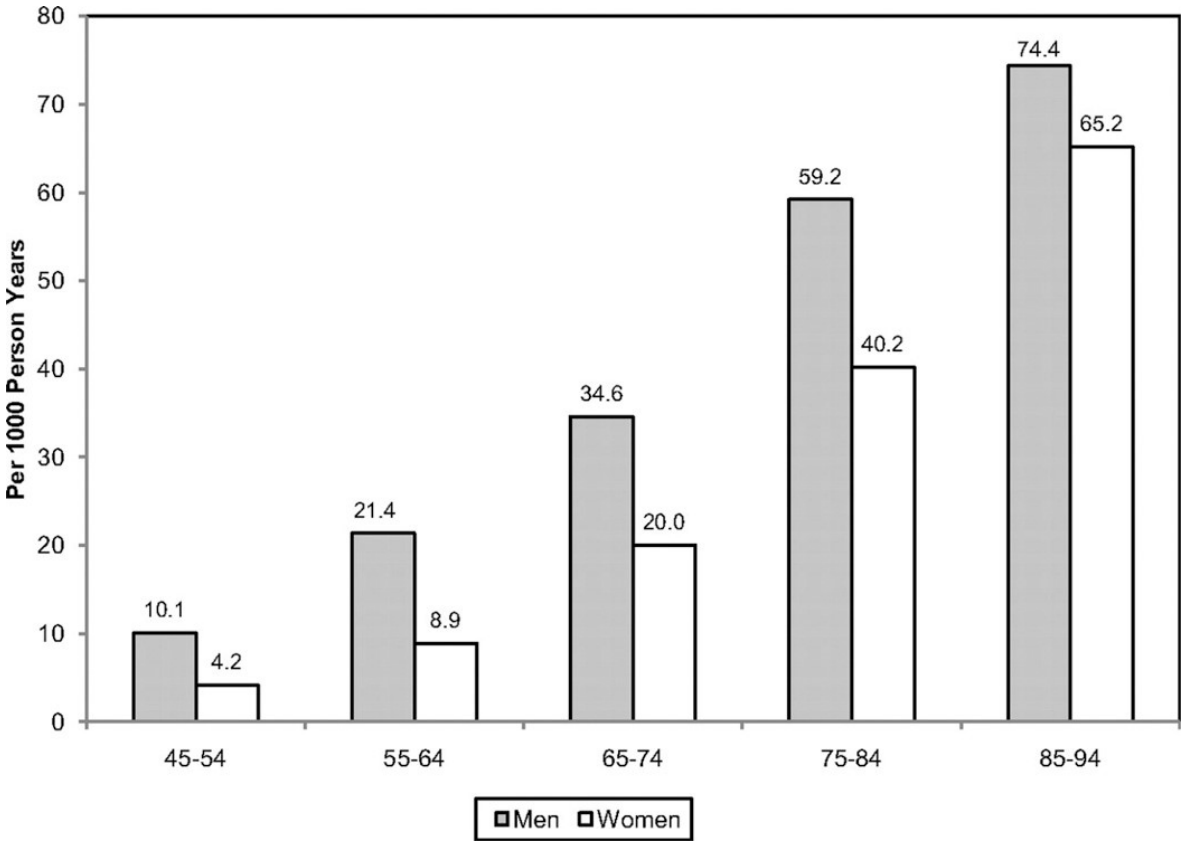



Figure 3 Annual number of adults having diagnosed heart attack or fatal coronary heart disease by age and sex.

As illustrated in Figure 3 the absolute number of annual CVD deaths among the female sex has exceeded that of the male sex since 1984. These data are often confused with CVD mortality rates, which, when adjusted for differences in age distribution, reveal that the CVD mortality rate is substantially higher in men than women. In 2007, the age-adjusted CVD death rate in men was 300 per 100 000 compared with 212 per 100 000 women. The 2007 CVD mortality rate in women represents a 43% reduction from the rate in 1997. From 1980 to 2000, the age-adjusted death rate for CHD fell from 263 to 134 per 100 000 women; during the same time period, the rate fell from 543 to 267 per 100 000 men.[70]

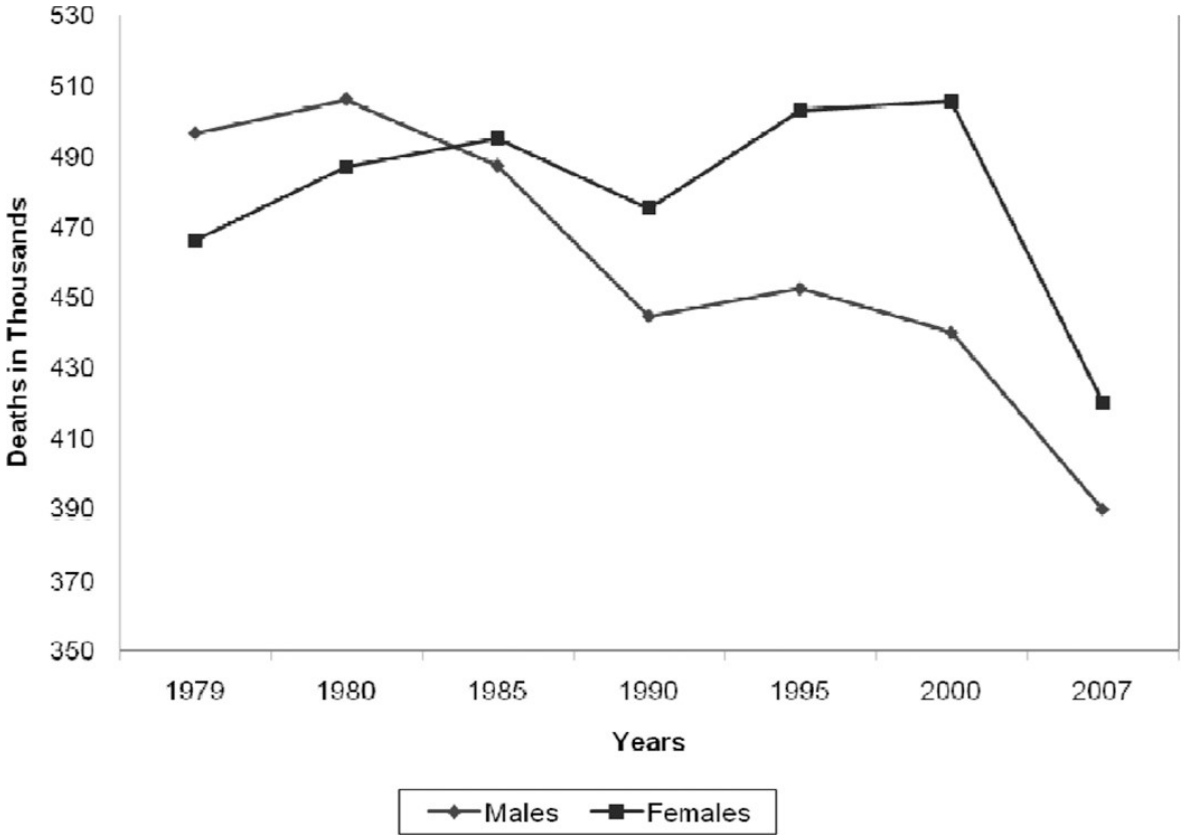



**Figure 4** Trends in the total annual number of deaths caused by cardiovascular disease according to gender

The prevalence of CVD in women varies according to racial/ethnic minority status. The prevalence of CVD among women ≥20 years of age is 47% among blacks, 34% among whites, and 31% among Mexican Americans; the prevalence of CHD is 7.6%, 5.8%, and 5.6%, respectively.[70] Asian women ≥18 years of age have the lowest prevalence of CHD (3.9%), according to the National Center for Health Statistics. The age-adjusted CHD death rate is highest among black women (122 per 100 000 compared with 94 per 100 000 in white women). The ominous trend for increasing rates of hypertension among black women is of particular concern because the increased risk for both CHD and stroke compared with white women could potentially widen the racial gap in CVD mortality. Dr Bernadine Healy first introduced the concept of the Yentl syndrome in 1991, suggesting gender bias in the management of CHD. There is ongoing debate as to whether women have a poorer prognosis after a myocardial infarction (MI) than men, and why. Is any observed difference explained by delay in women seeking care, healthcare provider delay in recognition and treatment, underlying differences in pathophysiology, more comorbidities, or older ages at time of presentation among women compared with men?[71]

Data over the past decade have shown that women have a higher 30-day mortality compared with men, and it is now recognized that the gender differences are largely explained by clinical differences at presentation. The higher mortality rate among women appears to be limited primarily to ST-segment–elevation MI. It has also been suggested that higher death rates may be restricted to younger women.[72] Although women with acute coronary syndromes may have similar benefits from antiplatelet pharmacotherapy as men, they are more likely to have bleeding problems, possibly as a result of excess dosing. Women experience greater morbidity and mortality than men after coronary artery bypass grafting; this disparity may reflect technical difficulties resulting from differences in body size, more microvascular disease, and different risk factor profiles. More recently, it has been shown that increasing use of off-pump coronary artery bypass grafting has narrowed the gender disparity in outcomes. Early studies that examined gender differences in outcomes after MI and revascularization may no longer be relevant owing to temporal trends in management and risk factor profiles.[73] Recent data from the National Registry of Myocardial Infarction showed that in-hospital mortality after an acute MI decreased more in women than in men between 1994 and 2006; the absolute reduction was 3 times larger in women than in men <55 years of age (2.7% versus 0.9%). This narrowing of the mortality gap was explained largely by greater improvements in risk factors in women than in men.[74] The classic risk factors for CVD are the same in women and men, but there are gender differences in the prevalence of risk factors. Although women and men overall have nearly equal percentages of hypertension (1 in 3 adults), data from the National Health and Nutrition Examination Survey (NHANES) showed that the prevalence of high blood pressure is greater in women >65 years of age. The highest rate of hypertension is among black women, 44%, and is increasing. The death rate caused by hypertension in 2007 was 37.0 per 100 000 for black women compared with 14.3 per 100 000 for white women. Diabetes mellitus is more prevalent among women than men ≥20 years of age (8.3% versus 7.2%).[75] Type II diabetes mellitus imparts a greater risk of CHD in women than men and is not explained by differences in risk factors, but rather by the more favorable survival rate of women (than men) without diabetes mellitus. The prevalence of physician-diagnosed diabetes mellitus is highest among non-Hispanic black (14.7%) and Mexican American (12.7%) women. On the basis of the NHANES data, the age-adjusted prevalence of the metabolic syndrome is highest among Mexican American women (40.6%), which is ≈22% higher than in Mexican American men. The prevalence of total cholesterol ≥240 mg/dL in 2008 for those ≥20 years of age was 16.2% among women

and 13.5% among men. In contrast, the percent of women with high-density lipoprotein cholesterol <40 mg/dL was 9.7% compared with 29.5% among men.[76]

Lifestyle risk factors also vary by gender, race, and ethnicity. Cigarette smoking has decreased overall in the United States, but remains more common among men than women (23.1% versus 18.1%). Non-Hispanic white women have a higher rate of smoking (20.7%) than black women (18.8%) and Hispanic women (9.4%). Age-adjusted rates of physical inactivity in 2009 were higher in women than men (34.5% versus 30.3%). Adverse trends in levels of physical activity (> 12 times a month) reveal a decline from 1988 to 2006 from 57% to 43% in men and from 49% to 43% in women.[77] The decreasing levels of physical activity parallel the rising rates of overweight and obesity in the United States. Two thirds of Americans are overweight or obese (72% of men and 64% of women) as defined by body mass index. Among women, non-Hispanic blacks and Mexican Americans are more likely to be obese than non-Hispanic whites (50% versus 45% versus 33%, respectively).[78] From 1999 to 2008, the increase in the prevalence of obesity was greater among men than women.[79] Full adherence to 3 heart-healthy lifestyle behaviors (smoking abstinence, physical activity, and fruit and vegetable intake) was nearly 50% higher among women than men without CHD in a 2000 sample of the US population. Overall adherence was low (<20%) for both women and men. These data suggest that population-wide approaches are needed to reduce the burden of CVD in both genders.[80]

**Closing the Gap in Preventive Care**

Adherence to guidelines for the prevention of CVD is suboptimal for women and men. The extent to which physician behaviors, patient behaviors, and environmental factors explain nonadherence is not established.[81] The limited systematic evaluation of provider performance in CVD preventive care makes it difficult to document gender differences in the delivery of care. Etiologic explanations for any observed gender differences in adherence to preventive recommendations are even more elusive. Most studies are conducted in select settings, use a variety of quality indicators, and report limited data on confounding or effect-modifying variables. Despite these research limitations, several themes consistently emerge regarding barriers to optimal preventive care. A fundamental barrier to implementation of prevention guidelines may be the guidelines themselves. Shaneyfelt et al evaluated the guidelines process and found that longer guidelines included more standards than shorter

guidelines but were more often ignored in practice.[82] Evidence-based recommendations were used more often than recommendations for practice not based on research evidence, and controversial recommendations were followed less often than those that were noncontroversial. A study of AHA/American College of Cardiology Guidelines showed that adherence was higher when the recommendations were supported by randomized, controlled clinical trials. Guidelines are more likely to be followed if they are easy to implement and come from a highly respected source.[83] The AHA has published 3 women-specific evidence-based guidelines between 2004 and 2011 for the prevention of CVD, but the extent to which these guidelines changed physician behavior or affected any gender gap in risk factor management is not known. The most recent AHA women's guideline 2011 update emphasized the importance of risk assessment to improve the quality of preventive care and highlighted challenges of available risk assessment tools: short-term focus, relevance of outcome measures (CVD versus CHD), and underestimation of risk in women. Further research is needed to determine whether improved risk assessment is associated with improved clinical outcomes.[84] Cabana et al evaluated 76 studies describing barriers to adherence to clinical practice guidelines; lack of awareness, lack of familiarity, lack of agreement, lack of self-efficacy, lack of outcome expectancy, and inertia of previous practice were recurring thematic barriers for following guidelines. It was suggested that AHA guidelines for the prevention of CVD in women are heterogeneous, and consequently there are different barriers to implementation of individual recommendations.[85] In a national AHA study of 500 randomly selected physicians, the most commonly cited barriers to implementation of CVD prevention guidelines were time, insurance coverage, and the patient. This study also revealed that physician assessment of CVD risk of the patient was the primary driver of quality preventive care. Gender disparities in treatment were explained largely by the provider's lower perceived CVD risk in women, despite a similar calculated risk compared with men. A subanalysis of this study suggested that solo practitioners and older physicians should be targeted to help promote the use of the guidelines. In a program designed to improve screening and management of CHD risk factors in women, internists and obstetricians/gynecologists were queried about barriers to primary prevention; physician time was perceived as a major barrier to the provision of preventive care.[86] The authors suggest that the current structure and reimbursement system for health care must be addressed if the gender gap in CVD preventive care is to be reduced. In a nationally representative sample of women, the most frequently cited barriers to heart health were confusion in the media (49%),

the belief that health is determined by a higher power (44%), and caretaking responsibilities (36%). Psychosocial factors may also contribute to nonadherence to preventive recommendations in women. For example, depression and social isolation have been linked to CVD risk and may be mediated by nonadherence to preventive recommendations, although there is a lack of clinical trials to document that treatment of psychosocial risk improves patient outcomes. The roles of body image and other psychological, social, and cultural factors as mediators of nonadherence deserve further study. Systems approaches to CVD prevention have the potential to improve outcomes and to reduce disparities. The Get With the Guidelines Quality Improvement Program has shown improved adherence to secondary prevention guidelines over time for both women and men, but the data are subject to selection bias and secular trends.[87]

### 1.5.3 The association between blood pressure and mortality in patients with heart failure.

Blood pressure is the force that pumps blood around the circulatory system. When blood flow is restricted or blocked completely, the heart muscle is starved of oxygen. This leads to a heart attack. During a heart attack, blood pressure can go up, down, or remain constant, depending on how the body responds.[88]

**Increase in blood pressure -** Blood pressure might rise during a heart attack because hormones, such as adrenaline, are released. These hormones are released when the "fight or flight" response is triggered at times of intense stress or danger. This automatic response might make the heart beat faster and stronger.[89]

**Decrease in blood pressure -** Blood pressure might drop if someone is having a heart attack because the heart is too weak to maintain it, as the muscle might have been damaged. The severe pain a person might feel during a heart attack could also trigger an automatic response, which might lead to decreased blood pressure and fainting.[90]

**Blood pressure and heart attacks -** If high blood pressure is left untreated, it could increase the risk of a heart attack. High blood pressure can be a measure of how hard the heart is having to work to pump blood around the body via the arteries, which is why doctors monitor it. A buildup of fat, cholesterol, and other substances within the arteries is called plaque. Over time,

plaque hardens, causing the arteries to narrow. This narrowing means it takes more pressure to push the blood through the network of tubes. When plaque breaks away from the wall of an artery, a blood clot is formed around the plaque. Heart attacks can happen because plaque or blood clots cause the blood supply to the heart to be disrupted or blocked. High blood pressure is not always a severe health problem, however. Even healthy people can experience raised blood pressure from time to time due to exercise or stress.[91]

**How is blood pressure measured?**

1. **Systolic blood pressure (SBP)** is the pressure in the arteries, as the heart pumps blood out to the body.
2. **Diastolic blood pressure (DBP)** is the pressure in the arteries between heart beats.

On blood pressure charts, the top number refers to the systolic pressure, while the number underneath refers to the diastolic pressure.

The association between low blood pressure and prognosis in the general population has been controversial, with some reports suggesting an increased mortality for patients with the lowest blood pressures. Whereas many standard heart failure therapies decrease blood pressure, the relationship between mortality and blood pressure in patients with heart failure has not been previously evaluated. We used the Digitalis Investigation Group trial database to evaluate retrospectively the relationship among systolic blood pressure (SBP), diastolic blood pressure (DBP), and survival among 5747 patients with New York Heart Association class II or III heart failure and left ventricular ejection fraction < or = 0.45. Cox proportional hazards models were used to identify covariates predictive of long-term mortality.[92]

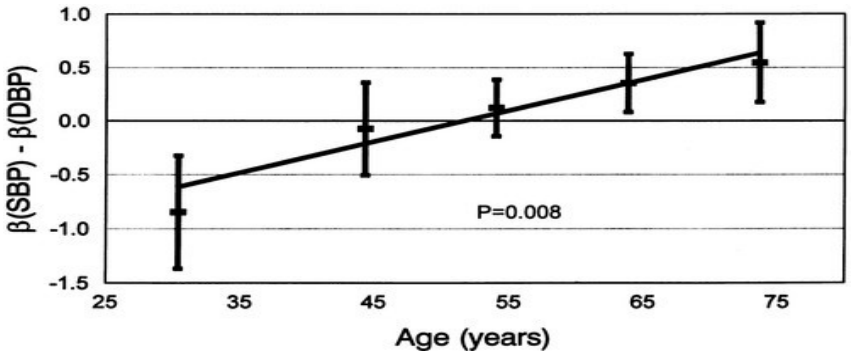



Figure 5 SBP and DBP wrt to Cardiovascular Disease Presence

**RESULTS:** The adjusted all-cause mortality rate during the entire study period for patients in the lowest SBP group (< 100 mm Hg) was 50% and was significantly higher than that of the reference group of patients with SBP of 130 to 139 mm Hg, which had a mortality rate of 32% (hazard ratio 1.65, 95% CI 1.25-2.17, P < .001). The relationship between SBP and mortality was significant (P < .001) and nonlinear (P = .009). The relationship between DBP and mortality was significant (P < .001), with the highest mortality seen in patients with DBP < 60 mm Hg. In patients with systolic dysfunction (left ventricular ejection fraction < or = 0.45) and New York Heart Association classes II and III symptoms, lower SBPs and DBPs were associated with greater mortality.[93]

### 1.5.4 Chest Pain and its risk factor to Cardiac arrest

Angina is chest pain or discomfort caused when your heart muscle doesn't get enough oxygen-rich blood. It may feel like pressure or squeezing in your chest. It is a symptom of an underlying heart problem, usually coronary heart disease (CHD).There are many types of angina, including microvascular angina, Prinzmetal's angina, stable angina, unstable angina and variant angina. This usually happens because one or more of the coronary arteries is narrowed or blocked, also called ischemia. Angina can also be a symptom of coronary microvascular disease (MVD). This is heart disease that affects the heart's smallest coronary arteries and is more likely to affect women than men. Coronary MVD also is called cardiac syndrome X and non-obstructive CHD. Learn more about angina in women.[94]

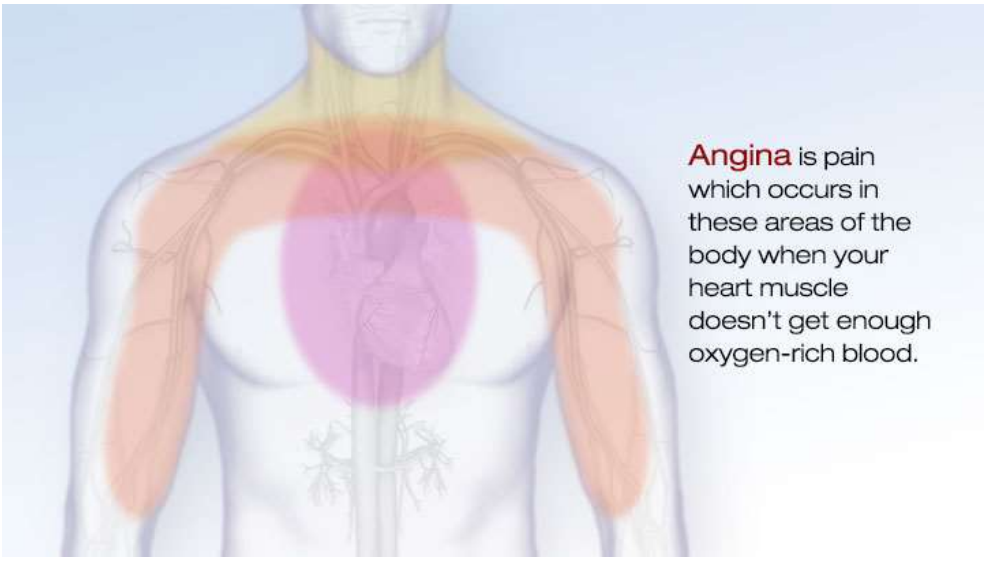



Figure 6 Pain Areas to be concerned about when having an Angina (Chest Pain)

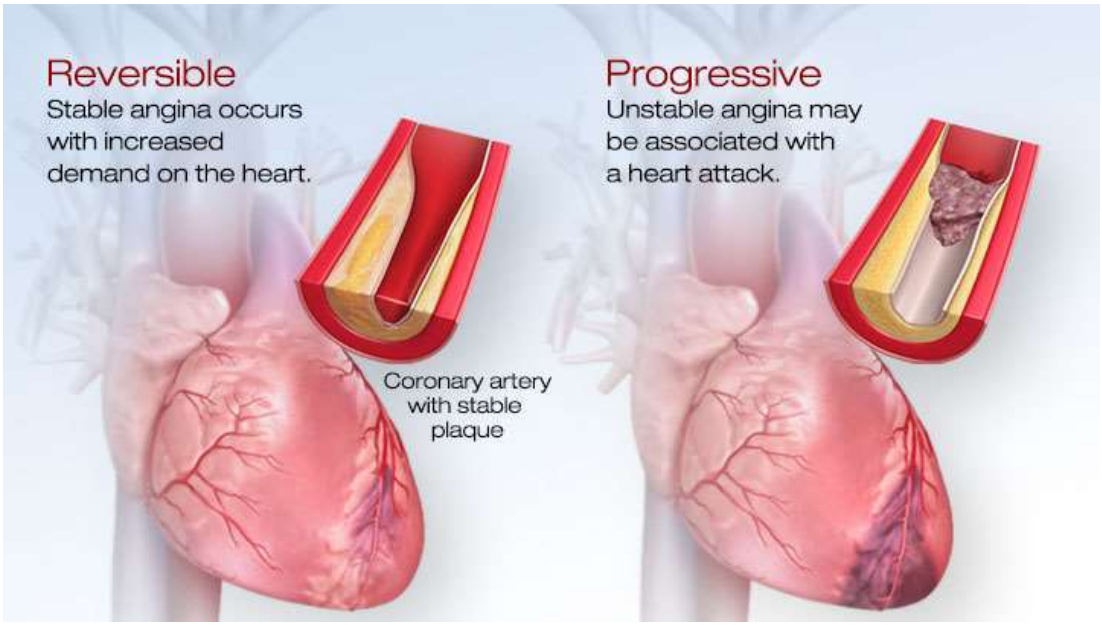



Figure 7 Reversible and Progressive Angina

**Types of Angina -** Knowing the types of angina and how they differ is important.

- Stable Angina / Angina Pectoris
- Unstable Angina
- Variant (Prinzmetal) Angina
- Microvascular Angina

**Diagnosis of Angina -** All chest pain should be checked out by a healthcare provider. If you have chest pain, your doctor will want to find out whether it's angina and if it is, whether the angina is stable or unstable. If it's unstable, you may need emergency medical treatment to try to prevent a heart attack.

Your doctor will most likely perform a physical exam, ask about your symptoms, and ask about your risk factors for and your family history of heart disease and other cardiovascular conditions.

### 1.5.5 Cholesterol and Heart Disease

Cholesterol helps your body build new cells, insulate nerves, and produce hormones. Normally, the liver makes all the cholesterol the body needs. But cholesterol also enters your body from food, such as animal-based foods like milk, eggs, and meat. Too much cholesterol in your body is a risk factor for heart disease.[91]

**How Does High Cholesterol Cause Heart Disease?**

When there is too much cholesterol in your blood, it builds up in the walls of your arteries, causing a process called atherosclerosis, a form of heart disease. The arteries become narrowed and blood flow to the heart muscle is slowed down or blocked. The blood carries oxygen to the heart, and if not enough blood and oxygen reach your heart, you may suffer chest pain. If the blood supply to a portion of the heart is completely cut off by a blockage, the result is a heart attack. There are two forms of cholesterol that many people are familiar with: Low-density lipoprotein (LDL or "bad" cholesterol) and high-density lipoprotein (HDL or "good" cholesterol.) These are the form in which cholesterol travels in the blood. LDL is the main source of artery-clogging plaque. HDL actually works to clear cholesterol from the blood. Triglycerides are another fat in our bloodstream. Research is now showing that high levels of triglycerides may also be linked to heart disease.[95]

**What Are the Symptoms of High Cholesterol?**

High cholesterol itself does not cause any symptoms, so many people are unaware that their cholesterol levels are too high. Therefore, it is important to find out what your cholesterol numbers are. Lowering cholesterol levels that are too high lessens the risk for developing heart

disease and reduces the chance of a heart attack or dying of heart disease, even if you already have it.[96]

**Do I need Treatment For High Cholesterol?**

Many health care providers recommend treating anyone with CVD with high-dose statin therapy. This includes those with coronary heart disease and who have had a stroke. For those who do not have CVD, treatment is determined by your individual risk for developing heart disease. That risk can be estimated using calculators which factor your age, sex, medical history, and other characteristics. If your risk is high (such as a 7.5 or 10 percent risk of developing CVD over 10 years), your doctor may start you on treatment preventively. They generally keep in mind your preferences towards taking medication in general. For those people whose risk is unclear, a coronary artery calcium score, which is a screening test looking for calcium (an indication of atherosclerosis) in the arteries, can help determine the need for statins. For both those who have CVD and those who do not, when the decision is made to start medication, the first choice is usually a statin.[97]

Other special groups who may need treatment:

- People with high triglyceride levels may benefit if they have other risk factors
- People with diabetes: are at high risk, and a ldl under 100 is recommended for most
- Older adults: a healthy, active older adult may benefit reduction you need, and prescribe a medication accordingly.

### 1.5.6 Fasting Glucose Level or Fasting Blood Sugar and the Risk of Heart Diseases

Both low glucose level and impaired fasting glucose should be considered as predictors of risk for stroke and coronary heart disease. The fasting glucose level associated with the lowest cardiovascular risk may be in a narrow range.[98] Diabetes is a well-established risk factor for cardiovascular disease (CVD) and all-cause mortality. Impaired fasting glucose (IFG), defined by the American Diabetes Association as having a fasting plasma glucose level of 100–125 mg/dL (5.6–7.0 mmol/L) or a 2-h value on the oral glucose tolerance test of 140–199 mg/dL (7.8–11.1 mmol/L) was associated with CVD risk in several studies.[99] The evidence is inconsistent, however, and the clinical relevance of IFG as a predictor of CVD is still unclear. In addition, the shape of the dose-response relationship between CVD risk and fasting glucose

level has not been well characterized across the full range of fasting blood glucose values. It is unclear whether there is an optimum fasting glucose level associated with the lowest level of CVD risk, or whether risk increases at very low fasting glucose levels.[95] Several studies have shown J-shape or U-shape relationships between fasting glucose levels and mortality. The Cancer Prevention Study (CPS) is a cohort study of >1.3 million adults designed to evaluate major risk factors for chronic diseases and mortality. The large sample size of this cohort facilitated detailed characterization of the dose-response relationship of fasting glucose level with the incidence of clinical CVD end points. In a large cohort of men and women, we found that fasting glucose level was associated with higher risk for major CVD outcomes, increasing from a level of ~90 mg/dL after controlling for other risk factors.[100] The dose-response curves showed progressive increments in the HRs from this value at both higher and lower levels; the increased risk was greatest for stroke. The patterns of association were similar in men and women, but the associations were stronger in women. Substantial evidence supports the biological plausibility of this finding. Experimental studies show that abnormal glucose metabolism impairs normal endothelial function, accelerates atherosclerotic plaque formation, and contributes to plaque rupture and thrombosis.[101] Epidemiological studies provide complementary evidence. In the Rotterdam Study, among elderly participants with a fasting blood glucose <110 mg/dL and without diabetes, those with higher blood glucose levels had higher levels of arterial stiffness.

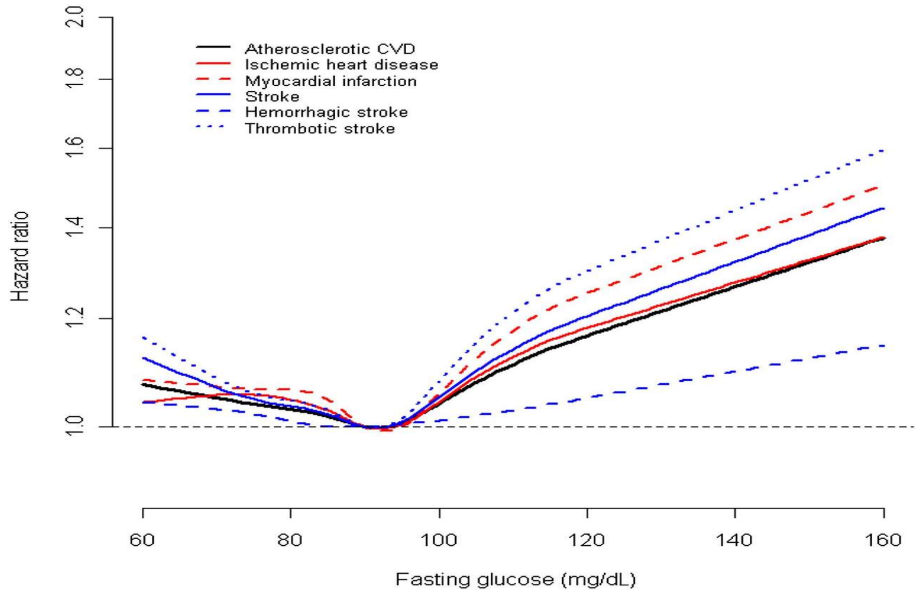



Figure 8 FBS analysis for Men

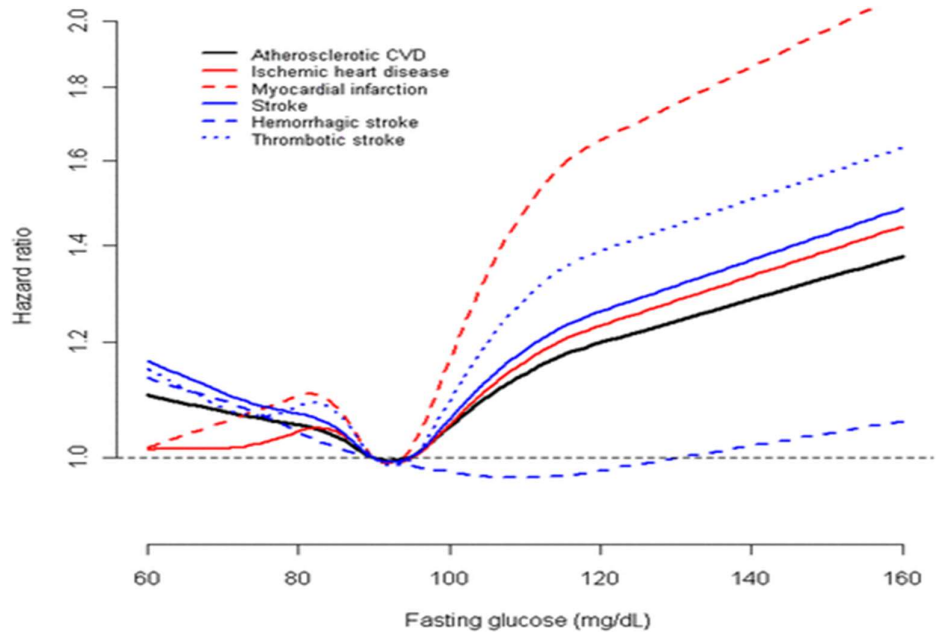



Figure 9 FBS analysis for Women

CATHAY study found that higher levels of glycemia (102–124 mg/dL) were associated with arterial endothelial dysfunction and intima-media thickening. In a biomarker study in Italy, a number of CVD biomarkers showed positive dose-response relationships with fasting glucose across three strata: <100; 100–109; and 110–125 mg/dL. Our study adds to the increasing evidence that IFG is an independent risk factor for incident CVD, including ischemic heart disease and stroke. In addition, the effects of other CVD risk factors may be enhanced by abnormal glucose metabolism.[102][103]

### 1.5.7 Electrocardiograph (ECG) Test for Heart Diseases

An electrocardiograph is the most common test for heart conditions. An electrocardiograph machine records your heart's rhythm onto paper through sticky electrodes which are placed on your chest, arms and legs. The recording will show if the heart muscle is damaged or short of oxygen. Specialized ECG tests:

- An exercise tolerance test (ETT) involves two ECG scans, one when you are exercising and one when you are resting. Some heart problems only appear when your heart needs to work harder. This test helps to show how your heart copes under stress.[104]

- A cardiac Holter monitoring test is used to identify any heart rhythm problems. For this test you wear a small, portable ECG machine for 24 or 48 hours and during this time your heart rate and rhythm are recorded.[105]
- Event monitoring is used to record your heartbeat when you experience symptoms such as dizziness, black outs, chest pain or palpitations. When you experience symptoms, you will need to press a button to start the recording.[106]

### 1.5.8 Cardiac Complications in Thalassemia Major

Thalassemia major is characterized by chronic ineffective erythropoiesis and anemia as its primary problems. These, in turn, produce physiologic adaptations in the cardiovascular system as well as pathologic/iatrogenic processes such as iron overload, splenectomy, nutritional deficiencies, chronic oxidative stress, and lung disease. This article discusses the pathophysiology of thalassemia as it relates to the cardiovascular system, the mechanisms and monitoring of iron cardiomyopathy, pulmonary hypertension, and vascular aging in thalassemia patients.[107]

1. **Chronic Anemia -** Patients with chronic anemia increase their cardiac output to maintain oxygen delivery, resulting in increased cardiac dimensions and heart rate. Anemic patients have larger hearts on CXR, echo, and MRI measurements than patients with normal hemoglobin levels, even without any other pathology. Thus, normative data generated from non-anemic patients is inappropriate for patients with hemoglobinopathies.[108] The larger cardiac dimensions, stroke volumes, and heart rates carry metabolic cost; chronically anemic patients have higher resting oxygen consumption and decreased reserves. Increased resting metabolism is also a source of increased oxidative stress, independent of the free-radical effects of iron. Patients with thalassemia have low or normal blood pressures, despite their increased cardiac output, because they have lower vascular resistance. Lower tonic vascular tone partially compensates for the increased chamber dimensions, but it leaves patients vulnerable to the endothelial toxicity of iron overload as well as making them less tolerant and responsive to the effects of afterload-reducing agents.
2. **Splenectomy -** Hypersplenism is relatively common in the thalassemia's and may necessitate spleen removal. Splenectomy may also be performed to lower blood

transfusion requirements. However, the spleen plays a critically important role in removing hematologic debris from the cardiovascular system. Phosphatidylserine positive platelets, platelet fragments, and red cell fragments are powerful pro-coagulants. They also inhibit nitric oxide, stimulate vasoconstricting substances such as endothelin and vasoconstricting prostaglandins, and produce endothelial proliferation[109]. The spleen also removes brittle senescent red cells from the circulation, suppressing intravascular hemolysis. Cell-free hemoglobin is a powerful oxidant and scavenger of nitric oxide. As a result, splenectomy is a strong risk factor for intravascular thrombosis and pulmonary hypertension.

3. **Iron Overload -** Patients with thalassemia develop iron overload through increased iron absorption and trans fusional therapy. Iron is toxic to all the endocrine glands that support the heart. Insulin-resistance and frank diabetes are relatively common. Hyperglycemia and insulin resistance are powerful oxidative stressors to the heart, worsening the effects of iron overload. Proper insulin sensitivity is also vital for efficient cardiac energy utilization. Iron may also poison the thyroid and parathyroid gland, impairing metabolism and calcium regulation respectively. Iron-mediated adrenal insufficiency may also manifest itself during metabolic stress. Deficiencies of growth hormone and the sex steroids impair cardiac function. Iron-mediated endocrine toxicity must be excluded in TM patients with cardiac failure.[110]
4. **Nutritional Deficiencies -** The hemoglobinopathies are a hypermetabolic state and inherently produce chronic oxidative stress. Broad-spectrum nutritional deficiencies are prevalent and may reinforce disease toxicity. Fat-soluble vitamin depletion is common, including vitamin A, D, E, and K, suggesting fat mal-absorption. The mechanisms and consequences are unknown. Vitamin D deficiency is associated with increased cardiac iron and decreased function, but causation has not been proven. Many trace metals are decreased, including selenium, zinc, and copper. B-vitamin levels are also low, particularly thiamine, riboflavin, and folate, most likely from consumption during ineffective erythropoiesis. Severe thiamine deficiency can have neurological and cardiac toxicity, whereas deficient riboflavin and folate may result in elevated homocysteine and endothelial toxicity. Carnitine deficiency is also relatively common in thalassemia and can impair cardiac function.[111]

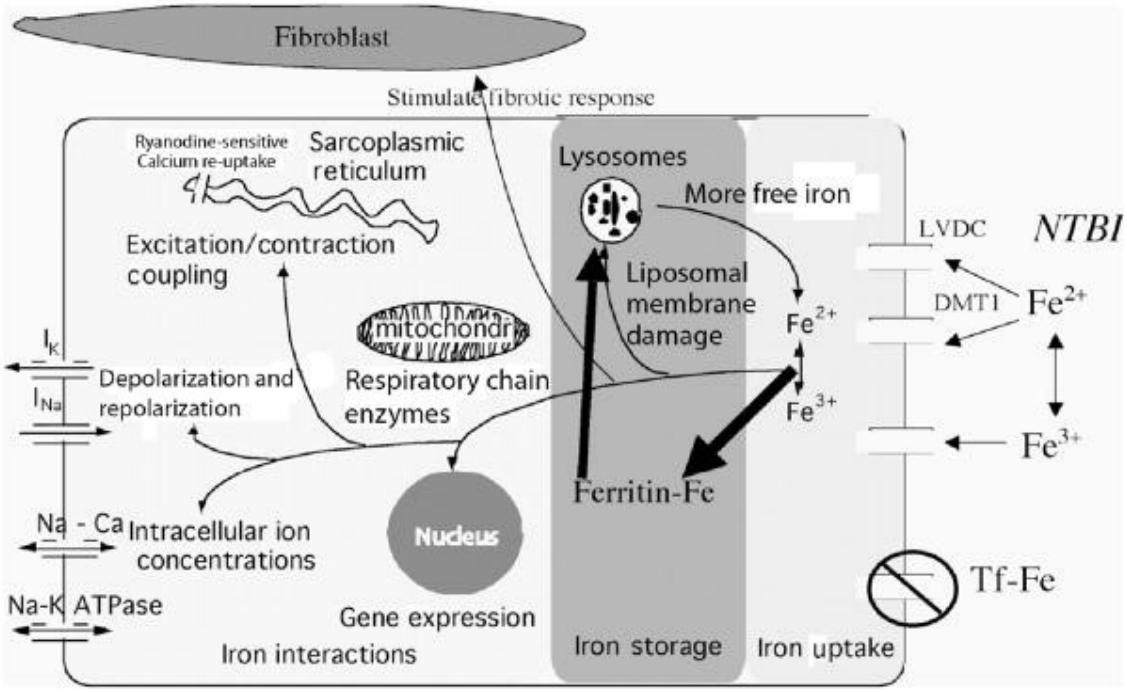



**Figure 10 Iron Cardiomyopathy -** represents the pathophysiology of iron cardiomyopathy, artificially divided into iron uptake, iron storage, and iron toxicity. The heart takes up physiologic amounts of iron through transferrin receptors, but this process is tightly regulated and does not lead to iron overload. When transferrin-binding capacity is exceeded, circulating low molecular weight non-transferrin-bound iron (NTBI) species appear. NTBI is oxidatively active and can enter through nonspecific, poor-regulated cation channels in the heart, leading to cardiac iron overload. Several channel mechanisms have been proposed, including L-type voltage-dependent calcium channels, but much more work is necessary to characterize cardiac iron-uptake processes.

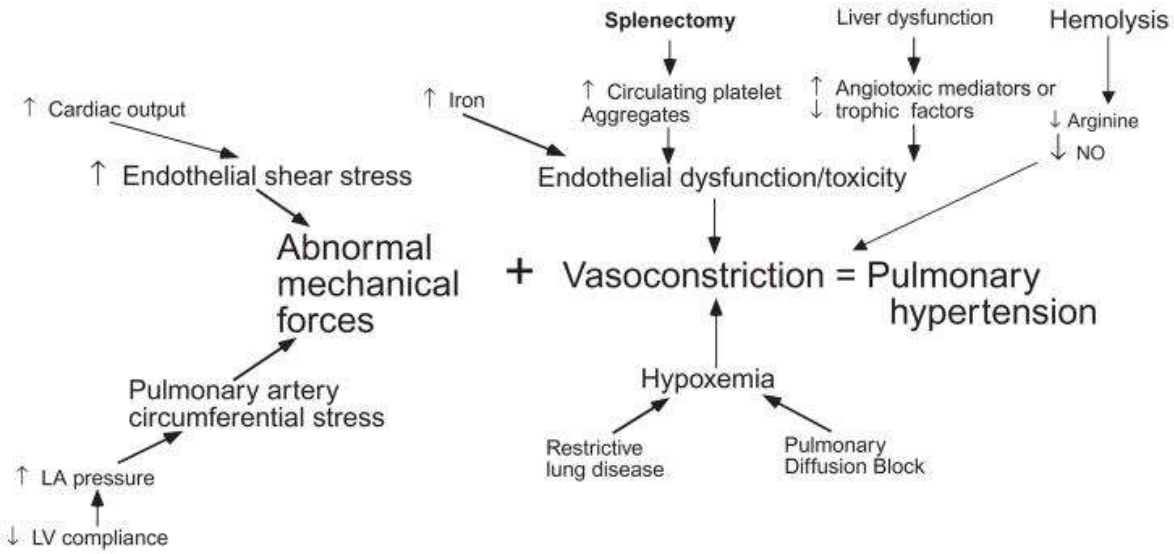



**Figure 11 Pulmonary Hypertension -** demonstrates the complex pathophysiology of pulmonary hypertension in thalassemia. Increased cardiac output and diastolic dysfunction cause abnormal loading of the pulmonary artery. Lung disease can exacerbate night-time hypoxia, a powerful stimulus for vasoconstriction. Iron, phosphatidylserine-expressing hematologic debris, free hemoglobin, and other circulating angiotrophic factors cause vasoconstriction and intimal proliferation.

# CHAPTER: 2 RELATED WORKS ON HEART DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS

## 2.1 Prediction system for heart disease using Naive Bayes and particle swarm

This section provides the basic concepts of classifier as Naive Bayes and feature subset selection method as PSO.

### 2.1.1 Particle swarm optimization (PSO)

PSO is an Evolutionary Computation technique is proposed by Kennedy et al. in 1995. PSO is motivated by social behaviors such as bird flocking and fish schooling. In PSO population swarm consists of "n" particles, and the position of each particle stands for the potential solution in D-dimensional space. The particles change its condition based on three aspects: To keep its inertia; To change the condition according to its most optimist position; To change the condition according to the swarm's most optimist position[30]. In PSO, a population are encoded as particles in the search space dimensionality D.PSO starts with the random initialization of a population of particles. Based on the best experience of one particle (pbest) and its neighboring particles (gbest), PSO searches for the optimal solution by updating the velocity and the position of each particle; PSO is used as feature subset selection method due to its advantages:

- Simple and easy to implement.
- Continuous optimization approach.

### 2.1.2 Naïve Bayes' Classifier

Naive Bayes classifiers are a family of simple probabilistic classifiers based by using Bayes theorem with strong (Naive) independence assumptions between the features. Naive Bayes classifiers are highly scalable by requiring several parameters linear for the number of features or predictors as variable in a learning problem.[112] It is the simplest and the fastest probabilistic classifier especially for the training phase.

**Feature selection -** It is a process of removing the irrelevant and redundant features from dataset based on evaluation criterion which is used to improve accuracy. There are two

approaches as individual evaluation and other one is subset evaluation. The process of feature selection is classified into three broad classes. One is filter and another one is wrapper and third one is embedded method based on how the feature selection is deployed by supervised learning algorithm. In this paper, they proposed a model which uses Naive Bayes as classifier and PSO as Feature subset selection measure for prediction of heart disease.[113]

**Proposed system -** In this section, we propose a methodology to improve the performance of Bayesian classifier for prediction of heart disease. Algorithm for our proposed model is shown below:

**Algorithm 1:** Heart disease prediction by using Bayes classifier and PSO.

Input: Heart disease dataset.

Output: Classify patient dataset into heart disease or not (normal).

Step 1: Read the dataset.

Step 2: Apply particle swarm optimization for feature selection.

Step 3: Remove the features with low value of PSO.

Step 4: Apply Naive Bayes classifier on relevant features. Step 5: Evaluate the performance of NB+PSO model.

The above algorithm divided into two sections, section 1 (step 2 and step 3) performs processing and feature subset selection. In section 2 (step 4 and step 5) Naive Bayes is applied on relevant features data and evaluate the performance in terms of accuracy.[114]

Accuracy= (No. of objects correctly classified/Total no. of objects in test set)

Cross validation technique used to split into training and testing data.

## 2.2 Predictive Data Mining for Medical Diagnosis: Heart Disease Prediction

The successful application of data mining in highly visible fields like e-business, marketing and retail has led to its application in other industries and sectors However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. Number of

experiment has been conducted to compare the performance of predictive data mining technique on the same dataset and the outcome reveals that Decision Tree outperforms and sometime Bayesian classification is having similar accuracy as of decision tree but other predictive methods like KNN, Neural Networks, Classification based on clustering are not performing well. The second conclusion is that the accuracy of the Decision Tree and Bayesian Classification further improves after applying genetic algorithm to reduce the actual data size to get the optimal subset of attribute enough for heart disease prediction.[9][115]

### 2.2.1 Data Mining in the Heart Disease Prediction.

Three different supervised machine learning algorithms i.e. Naive Bayes, K-NN, Decision List algorithm have been used for analyzing the dataset in. Tanagra tool is used to classify the data and the data is evaluated using 10-fold cross validation and the results are compared. **Tanagra** is a data mining suite build around graphical user interface algorithms. **Decision Tree** is a popular classifier which is simple and easy to implement. It requires no domain knowledge or parameter setting and can handle high dimensional data. The results obtained from Decision Trees are easier to read and interpret. The drill through feature to access detailed patients‟ profiles is only available in Decision Trees. **Naïve Bayes** is a statistical classifier which assumes no dependency between attributes. It attempts to maximize the posterior probability in determining the class. The advantage of using naive Bayes is that one can work with the naive Bayes model without using any Bayesian methods. Naive Bayes classifiers have works well in many complex real-world situations. The *k*-nearest neighbor's algorithm **(*k*-NN)** is a method for classifying objects based on closest training data in the feature space. k-NN is a type of instance-based learning. The *k*-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms. But the accuracy of the *k*-NN algorithm can be severely degraded by the presence of noisy or irrelevant features, or if the feature scales are not consistent with their importance. The experiment is performed using training data set consists of 3000 instances with 14 different attributes. The dataset is divided into two parts that is 70% of the data are used for training and 30% are used for testing.

### 2.2.2 Data Mining and Artificial Neural network

Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques, namely, Decision Trees, Naïve Bayes and Neural Network. is implemented in using .NET platform. IHDPS i s Web-based, user-friendly, scalable, reliable and expandable system. It can also answer complex "what if" queries which traditional decision support systems cannot. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. It enables significant knowledge, e.g. patterns, relationships between medical factors related to heart disease. As a **Data source** a total of 909 records with 15 medical attributes (factors) were obtained from the Cleveland Heart Disease database. Figure 1 lists the attributes. The records were split equally into two datasets: training dataset (455 records) and testing dataset (454 records). Initially, the data warehouse is pre- processed in order to make it suitable for the mining process. Once the preprocessing gets over, the heart disease warehouse is clustered with the aid of the K-means clustering algorithm, which will extract the data appropriate to heart attack from the warehouse. Consequently, the frequent patterns applicable to heart disease are mined with the aid of the MAFIA algorithm from the data extracted. In addition, the patterns vital to heart attack prediction are selected on basis of the computed significant weightage. The neural network is trained with the selected significant patterns for the effective prediction of heart attack. Multi-layer Perceptron Neural Network with Back-propagation is being used as the training algorithm. In feed-forward neural networks the neurons of the first layer forward their output to the neurons of the second layer, in a unidirectional fashion, which explains that the neurons are not received from the reverse direction.[116][24]

| Parameters | Weightage | |
|---|---|---|
| Male and Female | Age < 30 | 0.1 |
| | >30 to <50 | 0.3 |
| | Age>50 and Age <70 | 0.7 |
| | Age>70 | 0.8 |
| Smoking | Never | 0.1 |
| | Past | 0.3 |
| | Current | 0.6 |
| Overweight | Yes | 0.8 |
| | No | 0.1 |
| Alcohol Intake | Never | 0.1 |
| | Past | 0.3 |
| | Current | 0.6 |
| High salt diet | Yes | 0.9 |
| | No | 0.1 |
| High saturated fat diet | Yes | 0.9 |
| | No | 0.1 |
| Exercise | Never | 0.6 |
| | Regular | 0.1 |
| | High If age < 30 | 0.1 |
| | High If age > 50 | 0.6 |
| Sedentary Lifestyle/inactivity | Yes | 0.7 |
| | No | 0.1 |
| Hereditary | Yes | 0.7 |
| | No | 0.1 |
| Bad cholesterol | Very High >200 | 0.9 |
| | High 160 to 200 | 0.8 |
| | Normal <160 | 0.1 |
| Blood Pressure | Normal (130/89) | 0.1 |
| | Low (< 119/79) | 0.8 |
| | High (>200/160) | 0.9 |
| Blood sugar | High (>120&<400) | 0.5 |
| | Normal (>90&<120) | 0.1 |
| | Low ( <90) | 0.4 |
| Heart Rate | Low (< 60bpm) | 0.9 |
| | Normal (60 to 100) | 0.1 |
| | High (>100bpm) | 0.9 |

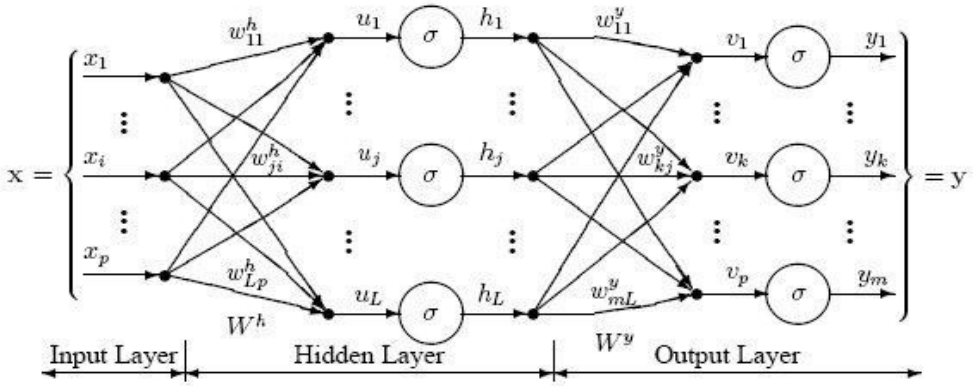



**Figure 12** The sample combinations of heart attack parameters for normal and risk level along with their values and weightages

### 2.2.3 Data Mining and Genetic Algorithm

Genetic algorithm has been used in, to reduce the actual data size to get the optimal subset of attributed sufficient for heart disease prediction. Classification is a supervised learning method to extract models describing important data classes or to predict future trends. Three classifiers e.g. Decision Tree, Naïve Bayes and Classification via clustering have been used to

diagnose the presence of heart disease in patients. **Classification via clustering:** Clustering is the process of grouping similar elements. This technique may be used as a preprocessing step before feeding the data to the classifying model. The attribute values need to be normalized before clustering to avoid high value attributes dominating the low value attributes. Further, classification is performed based on clustering. Experiments were conducted with Weka 3.6.0 tool. Data set of 909 records with 13 attributes. All attributes are made categorical and inconsistencies are resolved for simplicity. To enhance the prediction of classifiers, genetic search is incorporated. Observations exhibit that the Decision Tree data mining technique outperforms other two data mining techniques after incorporating feature subset selection but with high model construction time. Naïve Bayes performs consistently before and after reduction of attributes with the same model construction time.[117]

### 2.2.4 Association Rule Discovery

Association rules represent a promising technique to improve heart disease prediction. Unfortunately, when association rules are applied on a medical data set, they produce an extremely large number of rules. Most of such rules are medically irrelevant and the time required to find them can be impractical. In, four constraints were proposed to reduce the number of rules: item filtering, attribute grouping, maximum itemset size, and antecedent/consequent rule filtering. When association rules are applied on a medical data set, they produce an extremely large number of rules. Most of such rules are medically irrelevant and the time required to find them can be impractical. A more important issue is that, in general, association rules are mined on the entire data set without validation on an independent sample. To solve these limitations, the author has introduced an algorithm that uses search constraints to reduce the number of rules, searches for association rules on a training set, and finally validates them on an independent test set. Instead of using only Support and confidence, one more parameter i.e. lift have been used as the metrics to evaluate the medical significance and reliability of association rules. Medical doctors use sensitivity and specificity as two basic statistics to validate results. Sensitivity is defined as the probability of correctly identifying sick patients, whereas specificity is defined as the probability of correctly identifying healthy individuals. Lift was used together with confidence to understand sensitivity and specificity. To find predictive association rules in a medical data set the algorithm has three major steps. First, a medical data set with

categorical and numeric attributes is transformed into a transaction data set. Second, four constraints mentioned above are incorporated into the search process to find predictive association rules with medically relevant attribute combinations. Third, a train and test approach are used to validate association rules.[118][119]

### 2.2.5 Issues and Challenges

Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. The automation of the same would be highly beneficial. Clinical decisions are often made based on doctor's intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Data mining have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions.

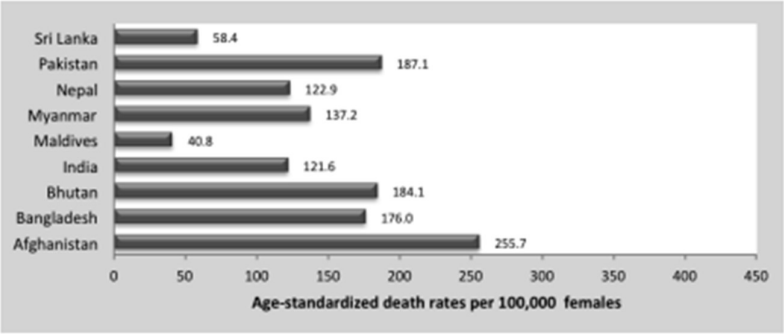
# CHAPTER 3: PRESENT WORK

## 3.1 About Heart Disease

Heart Diseases affect a large population in today's world, where the lifestyle is moved from active to comfort-oriented. We live in era of fast foods. Which build up cholesterol, diabetes and many more factors which in turn affects the heart in some way or the other. According to the World Health Organization Cardiovascular Diseases (CVD) or Heart Diseases cause more death than any other diseases globally[99]. The amount of data in medical sectors is quite large and computerized as well. They are not utilized or put to any use. This data if studied and analyzed could be put to good use like prediction of diseases or even prevent them. Diseases such as cancer can be detected, and the stage can also be predicted by training dataset with pictures of cancer cells. Similarly, heart disease can be predicted based on aspects like cholesterol, diabetes, heart rate etc. The prediction of heart diseases is a challenge and very risky. We observed that in some cases solutions of problems does not rely on a single method. It varies from situation to situation. It is also a challenge as most of the data are sparse or missing as they were not stored in the motive of analyzing. We therefore set out goal to finding which method would be best for predicting the diseases using data of four different hospitals from four different places. This is a comparative study on the efficiency of different data mining techniques such as Logical Regression, Random forest, K-Nearest Neighbors, Decision Tree in predicting heart diseases. The Data Mining techniques are analyzed, and the accuracy of prediction is noted for each method used. The result showed that heart diseases can be predicted with accuracy of above 90%. **Cardiovascular diseases** are the leading cause of death globally, resulted in 17.9 million deaths (32.1%) in 2015, up from 12.3 million (25.8%) in 1990. It is estimated that 90% of CVD is preventable. There are many risk factors for heart diseases that we will take a closer look at. The main objective of this study is to build a model that can **predict** the heart disease occurrence, based on a combination of features (risk factors) describing the disease. Different machine learning techniques will be implemented and compared upon standard performance metric such as accuracy. The dataset used for this study was taken from UCI machine learning repository[120].

## 3.2 Motivation

The rate of heart diseases is increasing at an exponential rate. The busy lifestyle of people in this era with all the fast food in the lunch break and getting back to sitting and working has pushed as over the edge. Along with this people today have a lack of exercise and are less active. For most of them recreation is just another movie in bed or anything technology based. Physical activities have reduced drastically. These factors boosted the rate of heart diseases to an unfortunately high percentage. In a developing country like ours the rate of heart diseases has the same effect. The annual mortality rate per 100,000 people from cardiovascular diseases in India has increased by 128.9% since 1990, an average of 5.6% a year[121].




Prediction of heart diseases is a difficult and risky task. Since it is directly dependent on people's' health, accuracy is a major factor. If not predicted accurately it can be disastrous. This research therefore focuses on the comparison of different data mining techniques to predict it. It shows the comparative analysis of the different methods. Cross validation error is used to compare the techniques. We choose Logical Regression, Random forest, K-Nearest Neighbors, Decision Tree as they are the most widely used techniques in determining diseases.

## 3.3 Proposal

In this project we have used 5 algorithms to find out the reasons of heart disease and create a model to get the maximum accuracy possible. For this we have used UCI dataset[122] where 4 databases: Cleveland, Hungary, Switzerland, and the VA Long Beach are already stored and the Creators are:

1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.

2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.

3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.

4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

And the Donor is: David W. Aha (aha '@' ics.uci.edu) (714) 856-8779. From this dataset we have used 5 widely used algorithms Logical Regression, Random forest, K-Nearest Neighbors and Decision Tree to create the model with the maximum accuracy possible. We have also explored precision score, recall score, F-score, false negative using confusion matrix for every algorithm used.[123]

## 3.4 Dataset Structure & Description

### 3.4.1 Importing libraries

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

%matplotlib inline

import os
print(os.listdir())

import warnings
warnings.filterwarnings('ignore')
```

```
['.config', 'heart.csv', 'sample_data']
```

### 3.4.2 Load data

```
data = pd.read_csv("heart.csv")
```

### 3.4.3 Check the type of the dataset

```
type(data)
```

```
pandas.core.frame.DataFrame
```

### 3.4.4 Check the Shape of the data

```
data.shape
```

```
(303, 14)
```

### 3.4.5 Check the top four columns of the dataset

```
data.head()
```

| | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 63 | 1 | 3 | 145 | 233 | 1 | 0 | 150 | 0 | 2.3 | 0 | 0 | 1 | 1 |
| 1 | 37 | 1 | 2 | 130 | 250 | 0 | 1 | 187 | 0 | 3.5 | 0 | 0 | 2 | 1 |
| 2 | 41 | 0 | 1 | 130 | 204 | 0 | 0 | 172 | 0 | 1.4 | 2 | 0 | 2 | 1 |
| 3 | 56 | 1 | 1 | 120 | 236 | 0 | 1 | 178 | 0 | 0.8 | 2 | 0 | 2 | 1 |
| 4 | 57 | 0 | 0 | 120 | 354 | 0 | 1 | 163 | 1 | 0.6 | 2 | 0 | 2 | 1 |

### 3.4.6: Dataset description

```
1 data.describe()
```

| | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 |
| mean | 54.366337 | 0.683168 | 0.966997 | 131.623762 | 246.264026 | 0.148515 | 0.528053 | 149.646865 | 0.326733 | 1.039604 | 1.399340 | 0.729373 | 2.313531 | 0.544554 |
| std | 9.082101 | 0.466011 | 1.032052 | 17.538143 | 51.830751 | 0.356198 | 0.525860 | 22.905161 | 0.469794 | 1.161075 | 0.616226 | 1.022606 | 0.612277 | 0.498835 |
| min | 29.000000 | 0.000000 | 0.000000 | 94.000000 | 126.000000 | 0.000000 | 0.000000 | 71.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 25% | 47.500000 | 0.000000 | 0.000000 | 120.000000 | 211.000000 | 0.000000 | 0.000000 | 133.500000 | 0.000000 | 0.000000 | 1.000000 | 0.000000 | 2.000000 | 0.000000 |
| 50% | 55.000000 | 1.000000 | 1.000000 | 130.000000 | 240.000000 | 0.000000 | 1.000000 | 153.000000 | 0.000000 | 0.800000 | 1.000000 | 0.000000 | 2.000000 | 1.000000 |
| 75% | 61.000000 | 1.000000 | 2.000000 | 140.000000 | 274.500000 | 0.000000 | 1.000000 | 166.000000 | 1.000000 | 1.600000 | 2.000000 | 1.000000 | 3.000000 | 1.000000 |
| max | 77.000000 | 1.000000 | 3.000000 | 200.000000 | 564.000000 | 1.000000 | 2.000000 | 202.000000 | 1.000000 | 6.200000 | 2.000000 | 4.000000 | 3.000000 | 1.000000 |

```
1 data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
age         303 non-null int64
sex         303 non-null int64
cp          303 non-null int64
trestbps    303 non-null int64
chol        303 non-null int64
fbs         303 non-null int64
restecg     303 non-null int64
thalach     303 non-null int64
exang       303 non-null int64
oldpeak     303 non-null float64
slope       303 non-null int64
ca          303 non-null int64
thal        303 non-null int64
target      303 non-null int64
dtypes: float64(1), int64(13)
memory usage: 33.2 KB
```

The dataset used in this project contains 14 variables. The independent variable that needs to be predicted, 'diagnosis', determines whether a person is healthy or suffer from heart disease. Experiments with the Cleveland database have concentrated on endeavors to distinguish disease presence (values 1, 2, 3, 4) from absence (value 0). There are several missing attribute values, distinguished with symbol '?'. The header row is missing in this dataset, so the column names have to be inserted manually.[124]

Features information:

- age - age in years
- sex - sex (1 = male; 0 = female)
- chest pain - chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic)
- blood pressure - resting blood pressure (in mm Hg on admission to the hospital)

- serum cholesterol - serum cholesterol in mg/dl
- fasting blood sugar - fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
- electrocardiographic - resting electrocardiographic results (0 = normal; 1 = having ST-T; 2 = hypertrophy)
- max heart rate - maximum heart rate achieved
- induced angina - exercise induced angina (1 = yes; 0 = no)
- ST depression - ST depression induced by exercise relative to rest
- slope - the slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 = down sloping)
- no of vessels - number of major vessels (0-3) colored by fluoroscopy
- thalassemia - 3 = normal; 6 = fixed defect; 7 = reversable defect
- diagnosis - the predicted attribute - diagnosis of heart disease (angiographic disease status) (Value 0 = < 50% diameter narrowing; Value 1 = > 50% diameter narrowing)

### 3.4.7 Types of features

**Categorical features** (Has two or more categories and each value in that feature can be categorized by them): **sex, chest pain**

**Ordinal features** (Variable having relative ordering or sorting between the values): **fasting blood sugar, electrocardiographic, induced angina, slope, no of vessels, thalassemia, diagnosis**

**Continuous features** (Variable taking values between any two points or between the minimum or maximum values in the feature column): **age, blood pressure, serum cholesterol, max heart rate, ST depression**

### 3.4.8 Some Random data columns

```
1 data.sample(5)
```

| | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **7** | 44 | 1 | 1 | 120 | 263 | 0 | 1 | 173 | 0 | 0.0 | 2 | 0 | 3 | 1 |
| **269** | 56 | 1 | 0 | 130 | 283 | 1 | 0 | 103 | 1 | 1.6 | 0 | 0 | 3 | 0 |
| **254** | 59 | 1 | 3 | 160 | 273 | 0 | 0 | 125 | 0 | 0.0 | 2 | 0 | 2 | 0 |
| **64** | 58 | 1 | 2 | 140 | 211 | 1 | 0 | 165 | 0 | 0.0 | 2 | 0 | 2 | 1 |
| **103** | 42 | 1 | 2 | 120 | 240 | 1 | 1 | 194 | 0 | 0.8 | 0 | 0 | 3 | 1 |

### 3.4.9 Check for missing Data

```
[10]  1 data.isnull().sum()
```

```
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
```

```
1 data.isnull().sum().sum()
```

```
0
```

No Data is missing, which is good.

### 3.4.10 Check the correlation with target data

```
print(data.corr()["target"].abs().sort_values(ascending=False))
```

```
target      1.000000
exang       0.436757
cp          0.433798
oldpeak     0.430696
thalach     0.421741
ca          0.391724
slope       0.345877
thal        0.344029
sex         0.280937
age         0.225439
trestbps    0.144931
restecg     0.137230
chol        0.085239
fbs         0.028046
Name: target, dtype: float64
```

**This shows that most columns are moderately correlated with target, but 'fbs' is very weakly correlated.**

## 3.5 Exploratory Data Analysis (EDA)

```
y = data["target"]

sns.countplot(y)


target_temp = data.target.value_counts()

print(target_temp)
```

```
1    165
0    138
Name: target, dtype: int64
```

(1 is who have Heart Disease and 0 is who don't have Heart Disease)

No. of Heart Disease patients is 165. No. of patients who don't have a heart disease is 138. [Which is a good balance of target data.]

```
1    165
0    138
Name: target, dtype: int64
```

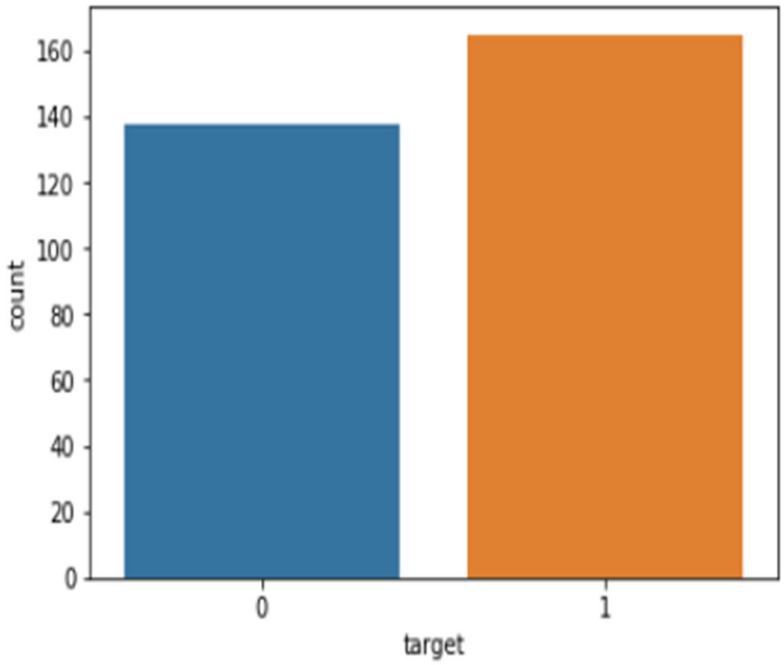



**Figure 13** Disease vs Non-Disease Analysis

### 3.5.1 Percentage of patient with or without heart problems in the given dataset

```
print("Percentage of patience without heart problems: "+str(round(target_temp[0]*100/303,2)))
print("Percentage of patience with heart problems: "+str(round(target_temp[1]*100/303,2)))
```

```
Percentage of patience without heart problems: 45.54
Percentage of patience with heart problems: 54.46
```

which is a good data distribution

### 3.5.2 Uniqueness of sex column –

Two sex types: 1 is male and 0 is female

```
data["sex"].unique()
```

```
array([1, 0])
```

### 3.5.3 Check the percentage and plot the graph

```
countFemale = len(data[data.sex == 0])
countMale = len(data[data.sex == 1])
print("Percentage of Female Patients:{:.2f}%".format((countFemale)/(len(data.sex))*100))
print("Percentage of Male Patients:{:.2f}%".format((countMale)/(len(data.sex))*100))
```

```
Percentage of Female Patients:31.68%
Percentage of Male Patients:68.32%
```

```
sns.barplot(data["sex"],y)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f77a8d6e828>
```

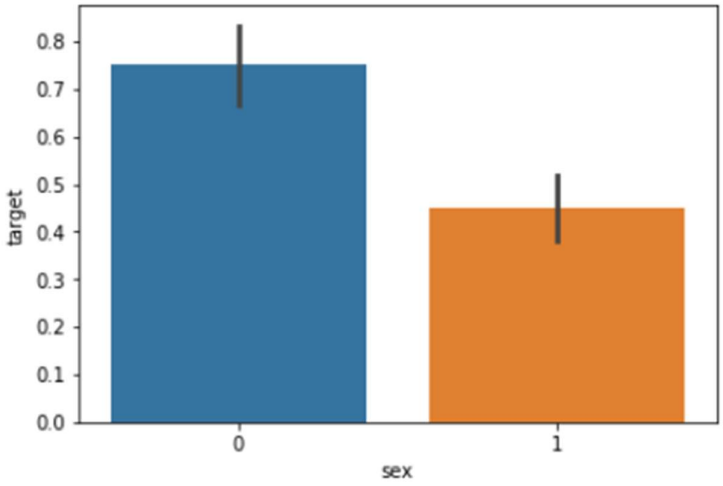



**Figure 13** Sex analysis

### 3.5.4 Heart Disease Frequency for Ages

```
pd.crosstab(data.age,data.target).plot(kind="bar",figsize=(20,6))
plt.title('Heart Disease Frequency for Ages')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.savefig('heartDiseaseAndAges.png')
plt.show()
```

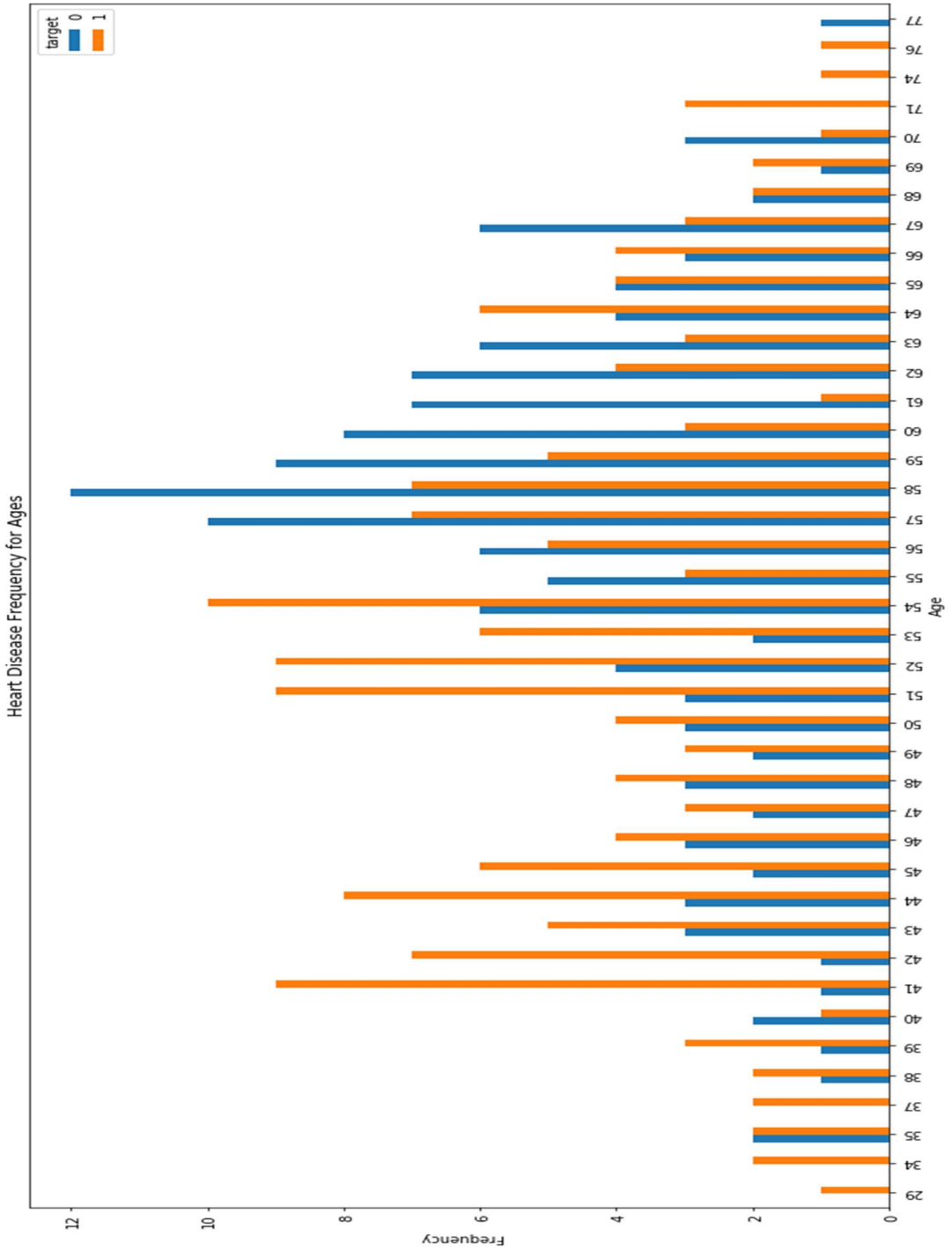



**Figure 14** Heart Disease Frequency for ages

### 3.5.5 Heart Disease Frequency for sex

```
pd.crosstab(data.sex,data.target).plot(kind="bar",figsize=(20,10),color=['blue','#AA1111' ])
plt.title('Heart Disease Frequency for Sex')
plt.xlabel('Sex (0 = Female, 1 = Male)')
plt.xticks(rotation=0)
plt.legend(["Don't have Disease", "Have Disease"])
plt.ylabel('Frequency')
plt.show()
```

**Where 1 is “male”,  0 is “female”.**

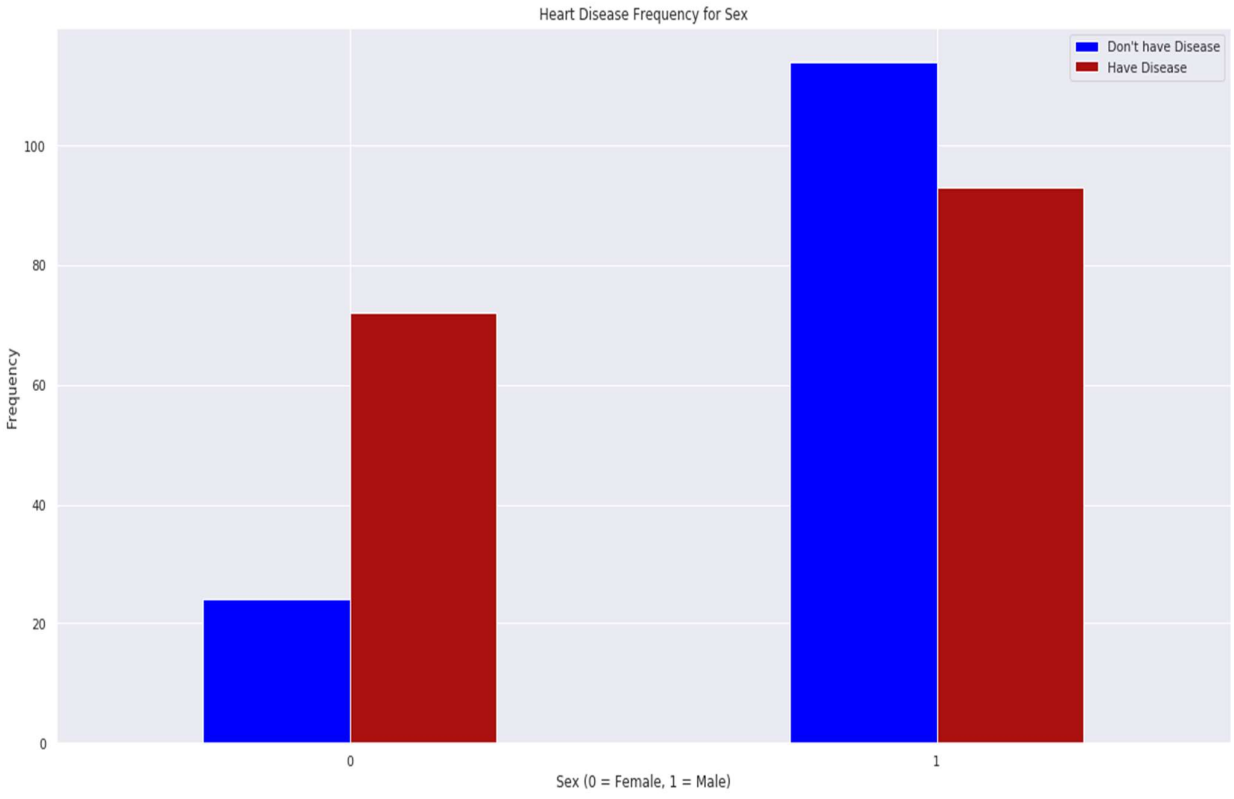



**Figure 15** Heart Disease Frequency for sex

### 3.5.6 Making the data column names easily recognizable

```
data.columns = ['age', 'sex', 'chest_pain_type', 'resting_blood_pressure', 'cholesterol', 'fasting_blood_sugar', 'rest_ecg', 'max_heart_rate_achieved',
       'exercise_induced_angina', 'st_depression', 'st_slope', 'num_major_vessels', 'thalassemia', 'target']
```

### 3.5.7 Checking out Male/Female Heart disease according to Fasting Blood Sugar

**Fasting blood sugar test:** A blood sample will be taken after an overnight fast. A fasting blood sugar level less than 100 mg/dL (5.6 mmol/L) is normal. A fasting blood sugar level from 100 to 125 mg/dL (5.6 to 6.9 mmol/L) is considered prediabetes. If it's 126 mg/dL (7 mmol/L) or higher on two separate tests, you have diabetes[125].

```
pd.crosstab(data.fasting_blood_sugar,data.target).plot(kind="bar",figsize=(20,10),color=['#4286f4','#f49242'])
plt.title("Heart disease according to FBS")
plt.xlabel('FBS- (Fasting Blood Sugar > 120 mg/dl) (1 = true; 0 = false)')
plt.xticks(rotation=90)
plt.legend(["Don't Have Disease", "Have Disease"])
plt.ylabel('Disease or not')
plt.show()
```

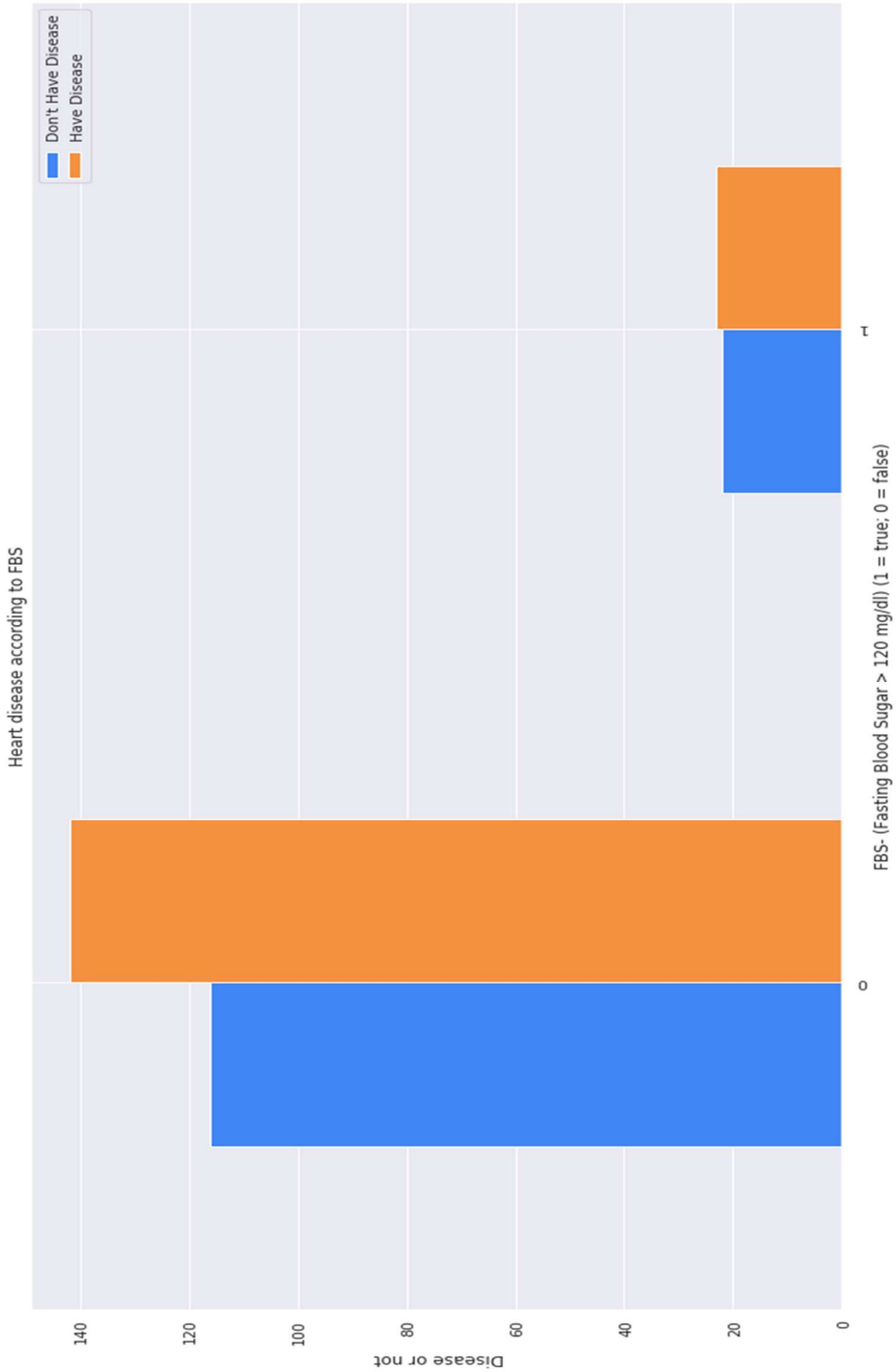


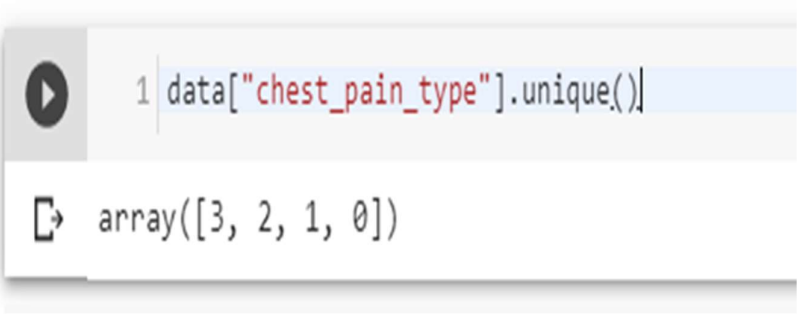

**Figure 16** Heart Disease according to FBS

### 3.5.8 Analyzing the chest pain

There are four types of Angina(chest pain)[126].

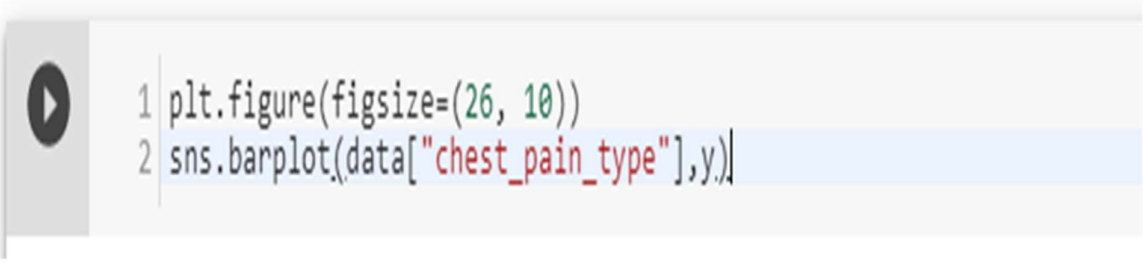
[Value 1: typical angina[122], Value 2: atypical angina[127], Value 3: non-anginal pain[123], Value 4: asymptomatic[128] ].

**Let's check how many types of chest pain (Angina) is present in our dataset**




4 types are present, 0, 1, 2, 3

**Let's plot chest pain types against target**



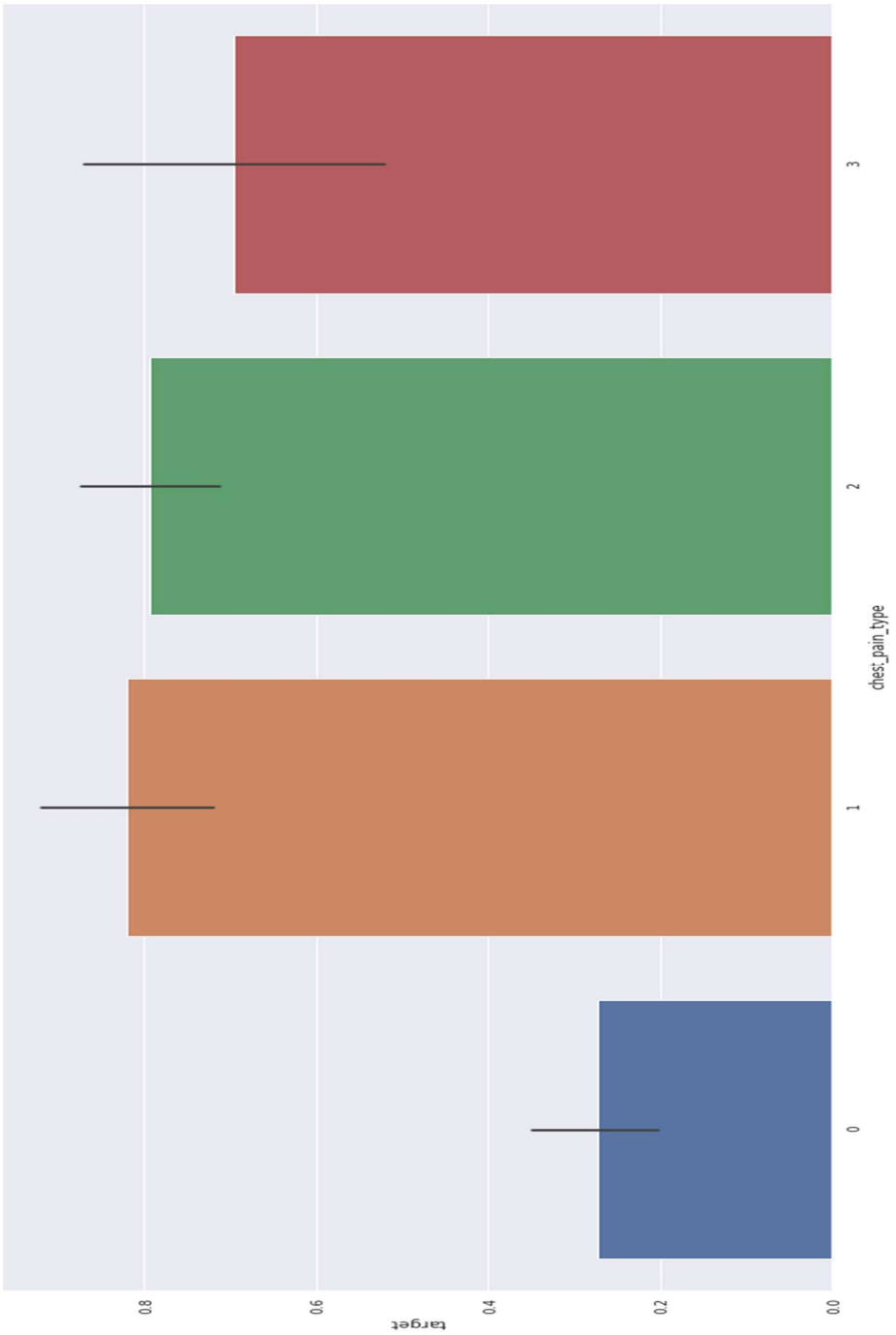



**Figure 17** Chest Pain Analysis

### 3.5.9 Analyzing the resting blood pressure

Resting blood pressure in mm Hg on admission to the hospital [129][124].

**Let's check out the unique resting blood pressures in our dataset**

```
1 data["resting_blood_pressure"].unique()
```

```
array([145, 130, 120, 140, 172, 150, 110, 135, 160, 105, 125, 142, 155,
       104, 138, 128, 108, 134, 122, 115, 118, 100, 124,  94, 112, 102,
       152, 101, 132, 148, 178, 129, 180, 136, 126, 106, 156, 170, 146,
       117, 200, 165, 174, 192, 144, 123, 154, 114, 164])
```

**Let's plot the resting blood pressure of our dataset against target**

```
1 plt.figure(figsize=(26, 10))
2 sns.barplot(data["resting_blood_pressure"],y)
```

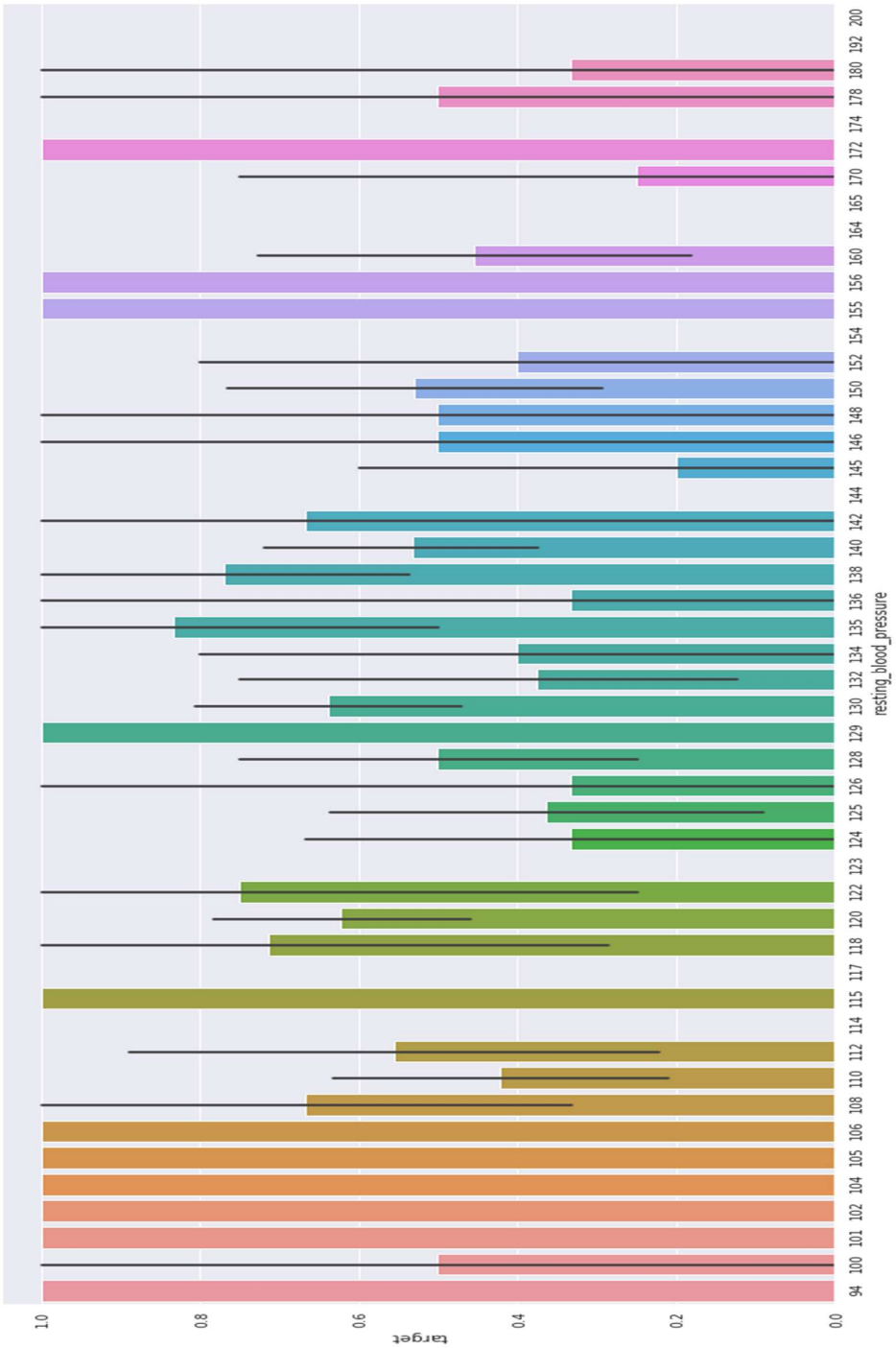



**Figure 18** Resting Blood Pressure Analysis

### 3.5.10 Analyzing the resting electrocardiographic measurement

0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria[130][131][132].

**Let's check uniqueness of our dataset**

```
1 data["rest_ecg"].unique()

array([0, 1, 2])
```

There are 3 types of resting ECG values are present: 0, 1, 2.

**Let's then plot the ECG values against target column**

```
1 plt.figure(figsize=(26, 10))
2 sns.barplot(data["rest_ecg"],y)

<matplotlib.axes._subplots.AxesSubplot at 0x7f840b654668>
```

**People with resting ECG value: 1 and 0 are much likely to have a heart disease than with the value 2 of resting ECG.**

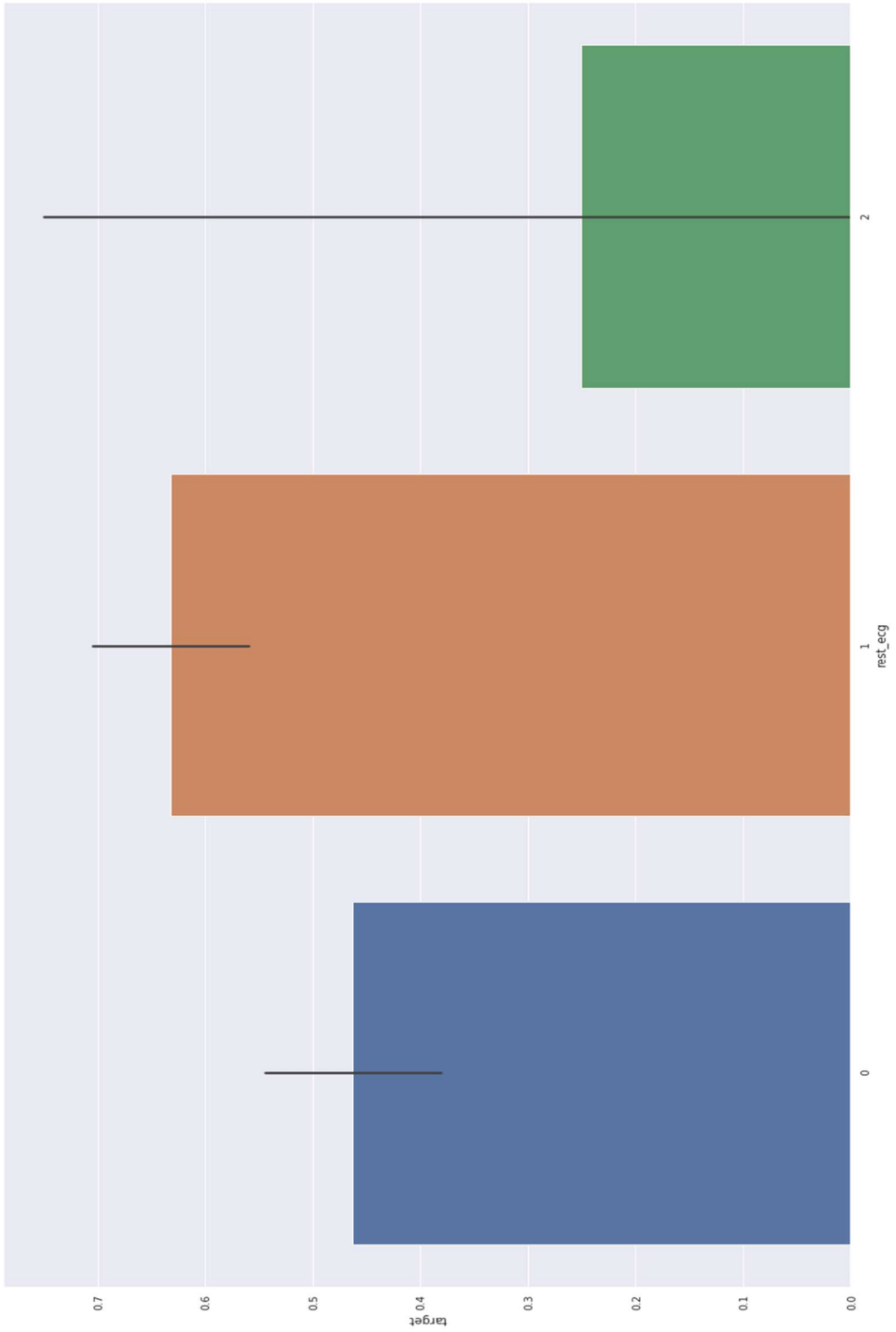



**Figure 19** ECG analysis

### 3.5.11 Analyzing Exercise Induced angina

1 means yes, and 0 means no.

```
data["exercise_induced_angina"].unique()
```

```
array([0, 1])
```

```
plt.figure(figsize=(26, 15))
sns.barplot(data["exercise_induced_angina"],y)
```

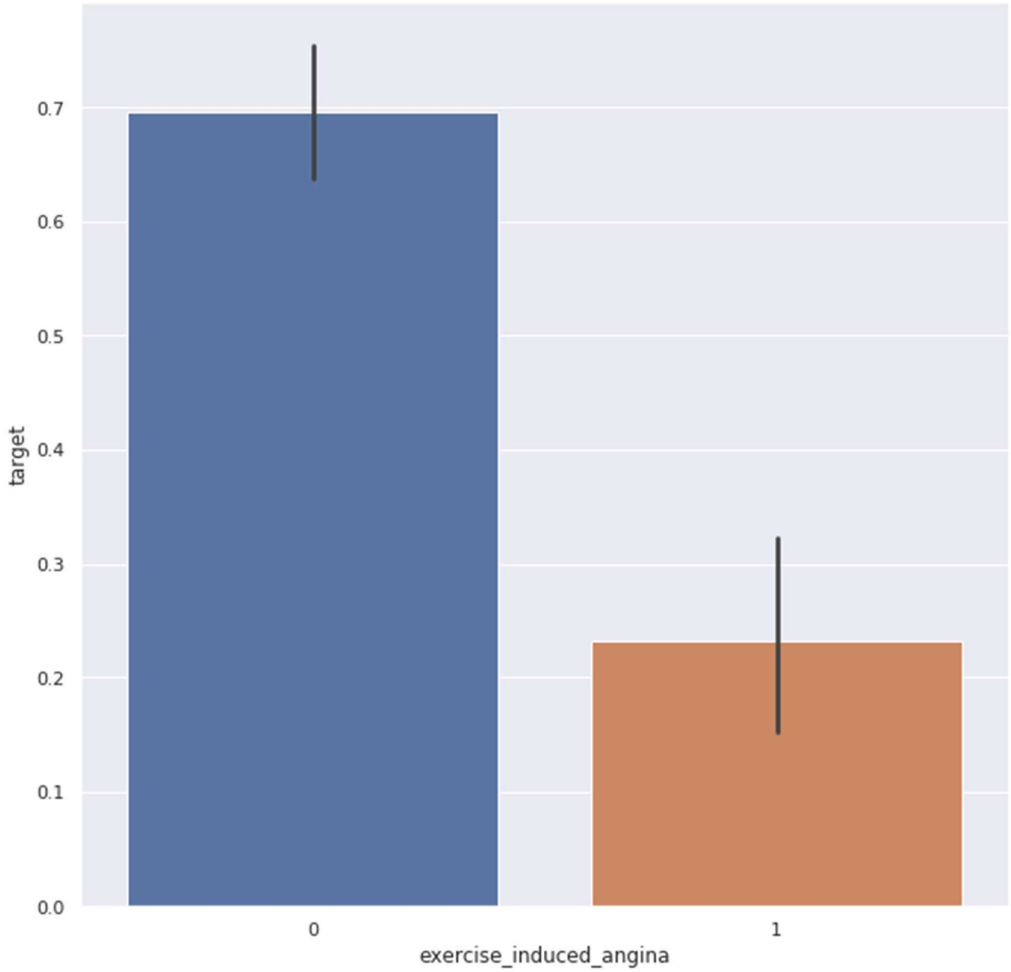



**Figure 20** Exercised Induced Angina Analysis

### 3.5.12 Slope of the peak exercise ST segment

The treadmill electrocardiogram (ECG) stress test is widely used to screen for obstructive coronary artery disease (CAD). The presence of ST segment changes, either depression or elevation, on the ECG during the treadmill test often suggests presence of CAD and warrants further management. We herein present three cases, with evidence of ischemia on the

treadmill ECG stress test. In addition, we discuss the use of the treadmill ECG stress test, including its indications, contraindications, reasons for termination and interpretation of the ST-segment changes, heart rate, as well as blood pressure responses to exercise[133].

```
1 data["st_slope"].unique()

array([0, 2, 1])
```

Value 1: upsloping, Value 2: flat, Value 3: down sloping.

```
1 plt.figure(figsize=(25, 10))
2 sns.barplot(data["st_slope"],y)

<matplotlib.axes._subplots.AxesSubplot at 0x7f840b4e64a8>
```

Slope '2' causes heart pain much more than Slope '0' and '1'.

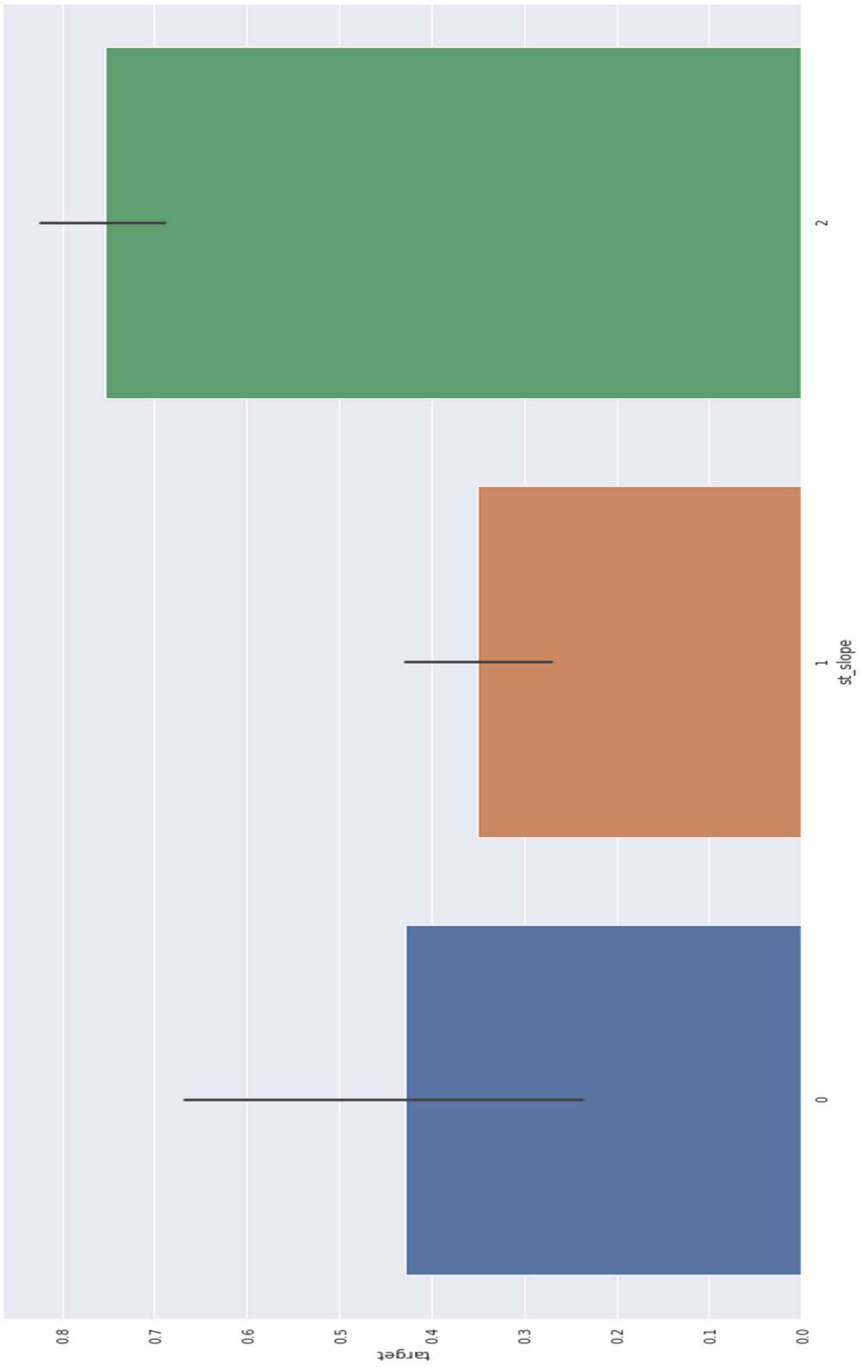



**Figure 21** ST Slope Analysis

### 3.5.13 Analyzing no. of major vessels colored by fluoroscopy

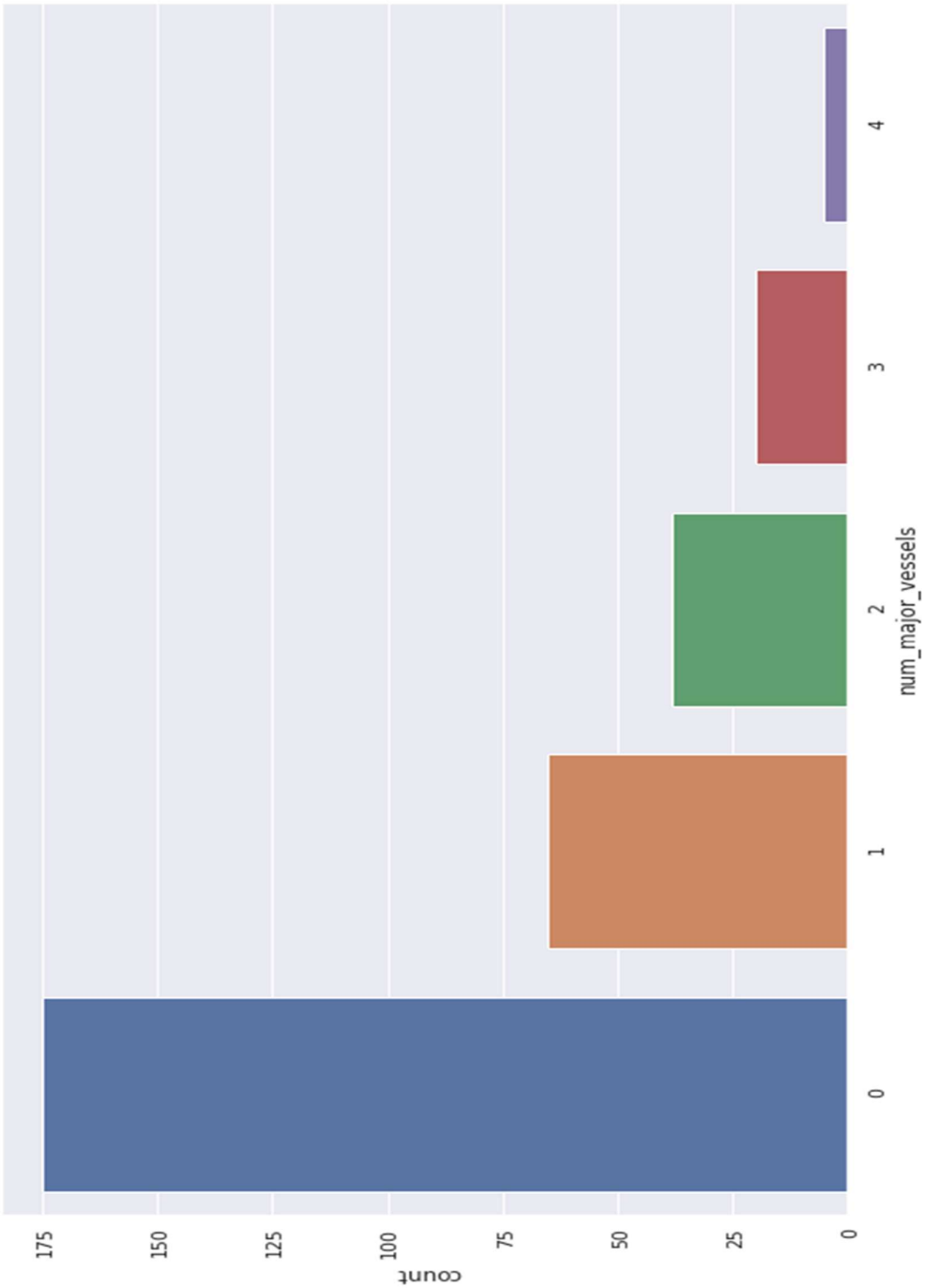



**Figure 22** Analyzing no. of major vessels colored by fluoroscopy

### 3.5.13.1 Comparing with target:

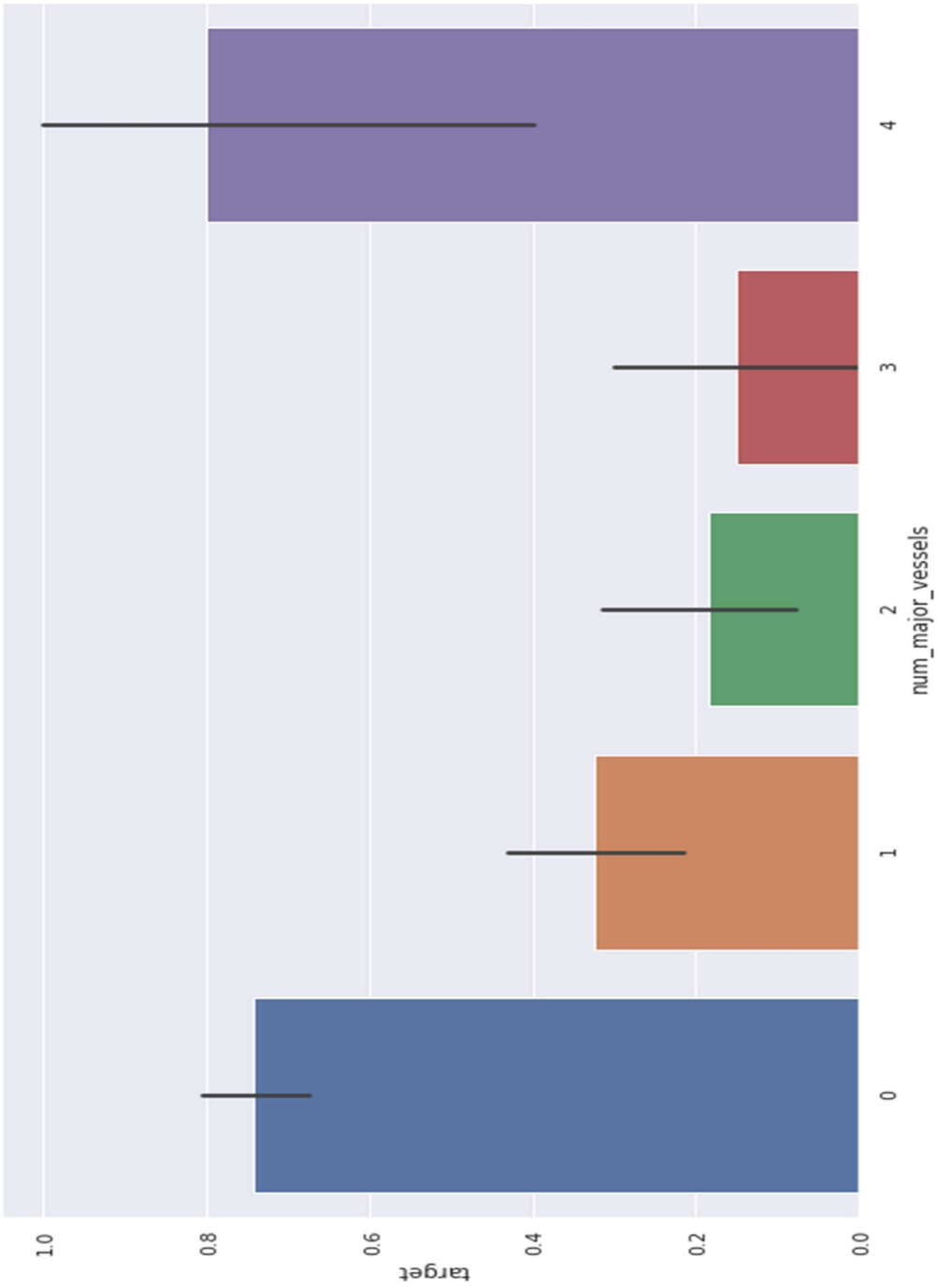



**Figure 23** Comparing with targets

### 3.5.14 Analyzing thalassemia

Four alpha-globin and two beta-globin protein chains make up hemoglobin. The two main types of thalassemia are alpha and beta.

**Alpha thalassemia -** In alpha thalassemia, the hemoglobin does not produce enough alpha protein. To make alpha-globin protein chains we need four genes, two on each chromosome 16. We get two from each parent. If one or more of these genes is missing, alpha thalassemia will result. The severity of thalassemia depends on how many genes are faulty, or mutated.

- **One faulty gene**: The patient has no symptoms. A healthy person who has a child with symptoms of thalassemia is a carrier. This type is known as alpha thalassemia minima.

- **Two faulty genes**: The patient has mild anemia. It is known as alpha thalassemia minor.

- **Three faulty genes**: The patient has hemoglobin H disease, a type of chronic anemia. They will need regular blood transfusions throughout their life.

- **Four faulty genes**: Alpha thalassemia major is the most severe form of alpha thalassemia. It is known to cause hydrops fetalis, a serious condition in which fluid accumulates in parts of the fetus' body. A fetus with four mutated genes cannot produce normal hemoglobin and is unlikely to survive, even with blood transfusions. Alpha thalassemia is common in southern China, Southeast Asia, India, the Middle East, and Africa.

**Beta Thalassemia -** We need two globin genes to make beta-globin chains, one from each parent. If one or both genes are faulty, beta thalassemia will occur.
Severity depends on how many genes are mutated.

- **One faulty gene**: This is called beta thalassemia minor.

- **Two faulty genes**: There may be moderate or severe symptoms. This is known as thalassemia major. It used to be called Colley's anemia.

Beta thalassemia is more common among people of Mediterranean ancestry. Prevalence is higher in North Africa, West Asia, and the Maldives Islands. So, we'll mainly work with

Alpha Thalassemia, And value 0 is one faulty gene, Value 1 is Two faulty genes, Value 2 is Three faulty genes, Value 3 is Four faulty genes.

**Thalassemia distribution**

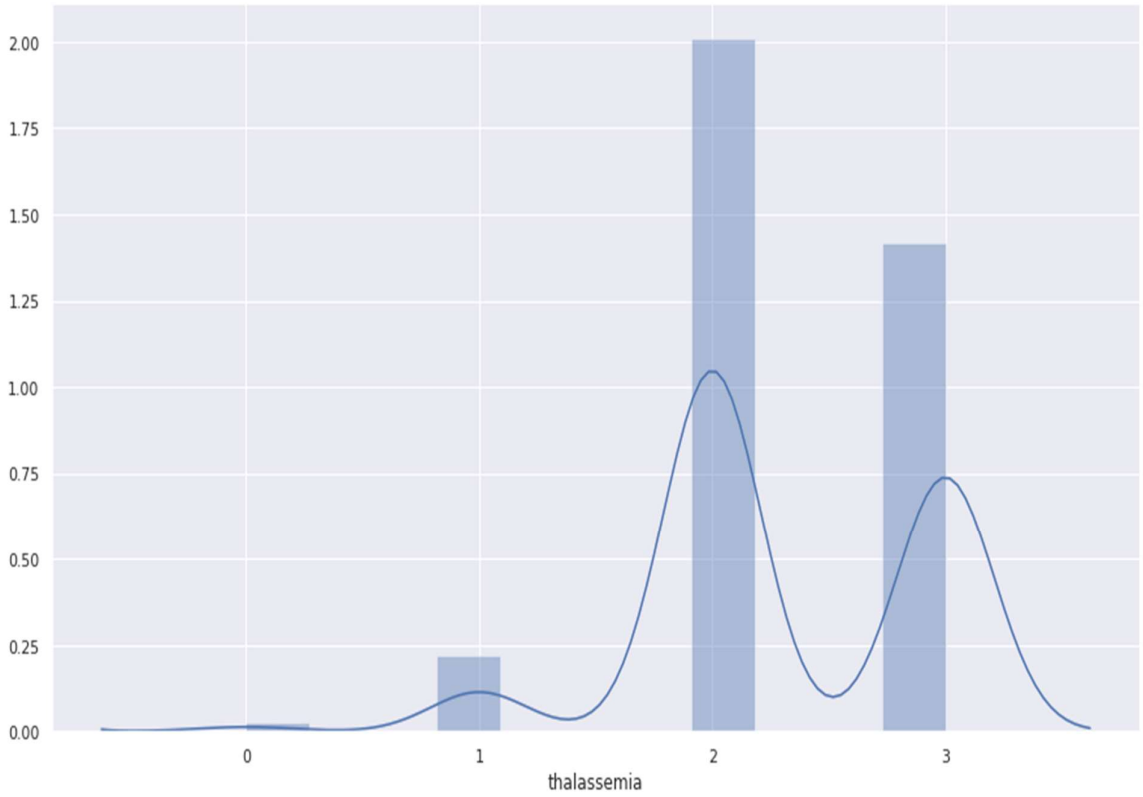



**Figure 24** Thalassemia distribution

**Against target:**

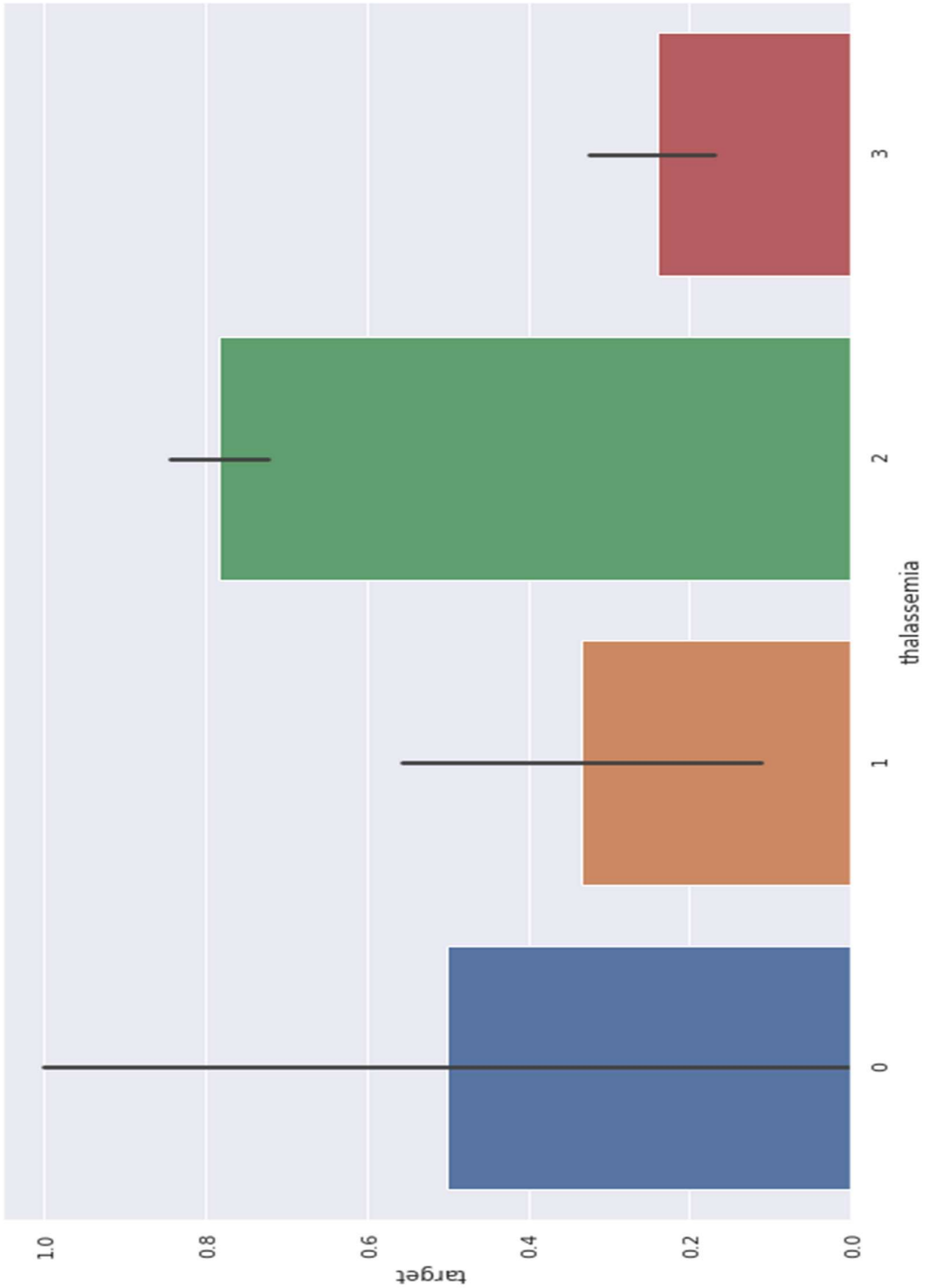



**Figure 25** Thalassemia against target

### 3.5.15 Thalassemia vs cholesterol

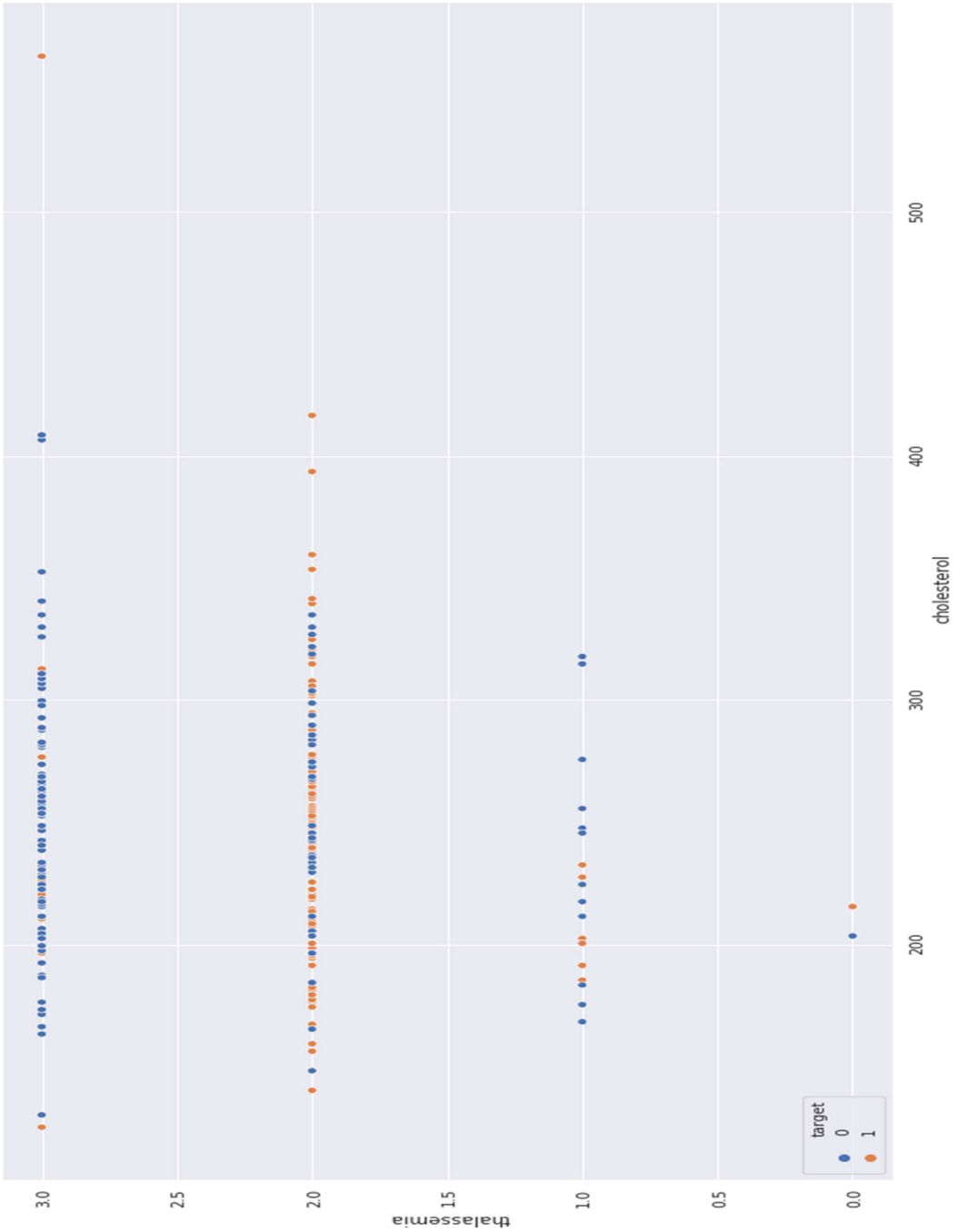



**Figure 26** Thalassemia vs cholesterol

## Thalassemia vs resting blood pressure


**Figure 27** Thalassemia vs resting blood pressure

## 3.6 Correlation Matrix

Correlation analysis is a method of statistical evaluation used to study the strength of a relationship between two, numerically measured, continuous variables (e.g. height and weight)

Store numeric variables in cname variable

```
# store numeric variables in cnames
cnames=['age','resting_blood_pressure','cholesterol','max_heart_rate_achieved','st_depression','num_major_vessels']
```

Let's plot the Correlation plot

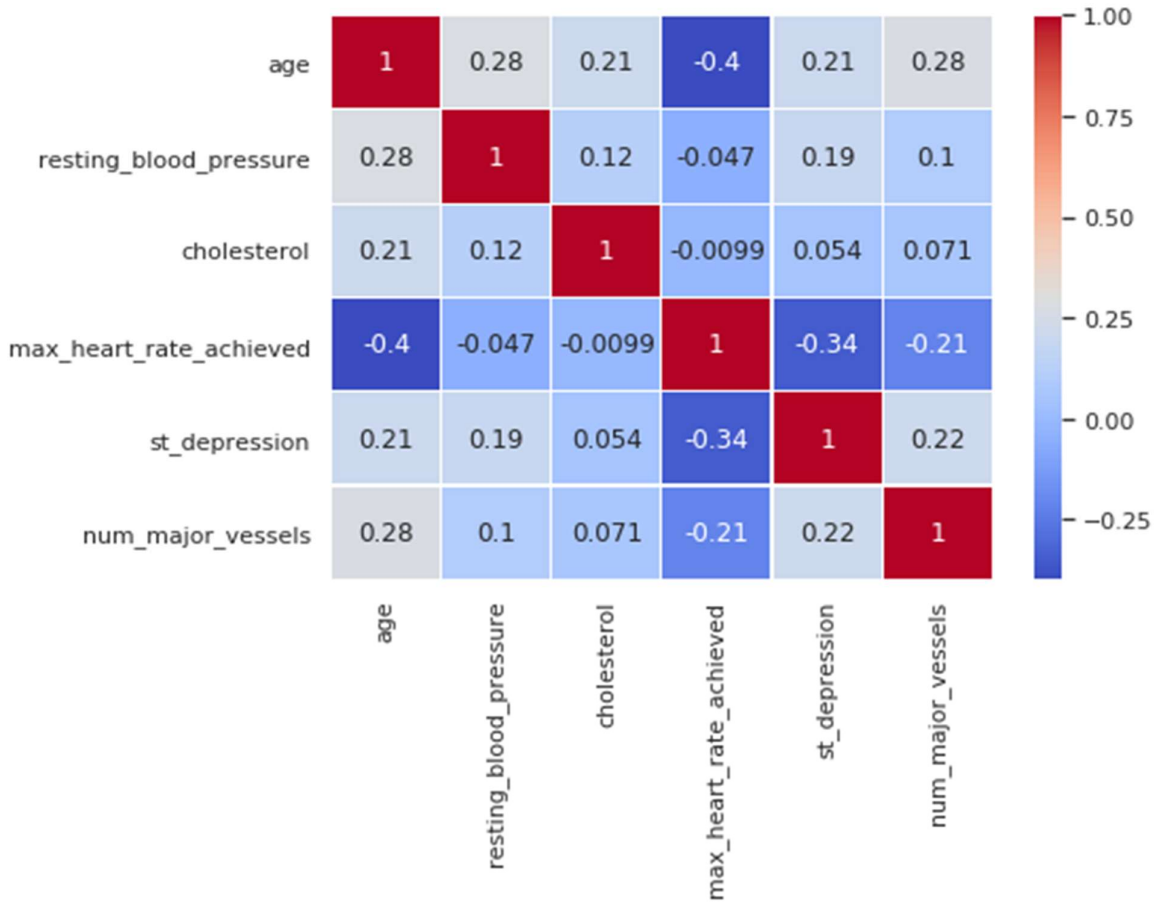



**Figure 28** Correlation Matrix

**Correlation Analysis:**

| | age | resting_blood_pressure | cholesterol | max_heart_rate_achieved | st_depression | num_major_vessels |
|---|---|---|---|---|---|---|
| **0** | 63 | 145 | 233 | 150 | 2.3 | 0 |
| **1** | 37 | 130 | 250 | 187 | 3.5 | 0 |
| **2** | 41 | 130 | 204 | 172 | 1.4 | 0 |
| **3** | 56 | 120 | 236 | 178 | 0.8 | 0 |
| **4** | 57 | 120 | 354 | 163 | 0.6 | 0 |
| **5** | 57 | 140 | 192 | 148 | 0.4 | 0 |
| **6** | 56 | 140 | 294 | 153 | 1.3 | 0 |
| **7** | 44 | 120 | 263 | 173 | 0.0 | 0 |
| **8** | 52 | 172 | 199 | 162 | 0.5 | 0 |
| **9** | 57 | 150 | 168 | 174 | 1.6 | 0 |
| **10** | 54 | 140 | 239 | 160 | 1.2 | 0 |
| ... | ... | ... | ... | ... | ... | ... |
| **291** | 58 | 114 | 318 | 140 | 4.4 | 3 |
| **292** | 58 | 170 | 225 | 146 | 2.8 | 2 |
| **293** | 67 | 152 | 212 | 150 | 0.8 | 0 |
| **294** | 44 | 120 | 169 | 144 | 2.8 | 0 |
| **295** | 63 | 140 | 187 | 144 | 4.0 | 2 |
| **296** | 63 | 124 | 197 | 136 | 0.0 | 0 |
| **297** | 59 | 164 | 176 | 90 | 1.0 | 2 |
| **298** | 57 | 140 | 241 | 123 | 0.2 | 0 |
| **299** | 45 | 110 | 264 | 132 | 1.2 | 0 |
| **300** | 68 | 144 | 193 | 141 | 3.4 | 2 |
| **301** | 57 | 130 | 131 | 115 | 1.2 | 1 |
| **302** | 57 | 130 | 236 | 174 | 0.0 | 1 |

303 rows × 6 columns

There is no single feature that has a very high correlation with our target value. Also, some of the features have a negative correlation with the target value and some have positive.

## 3.7 Data Preparation

Total Among 303 data's randomly 242 are chosen for Training and 61 are chosen for Testing**.**

```
[128]  1 X_train.shape
```

```
(242, 13)
```

```
[129]  1 X_test.shape
```

```
(61, 13)
```

```
[130]  1 Y_train.shape
```

```
(242,)
```

```
[131]  1 Y_test.shape
```

```
(61,)
```

## 3.8 Modelling and predicting with Machine Learning

The main goal of the entire project is to predict heart disease occurrence with the highest accuracy. In order to achieve this, we will test several classification algorithms. This section includes all results obtained from the study and introduces the best performer according to accuracy metric. We have chosen several algorithms typical for solving supervised learning problems throughout classification methods.

Modelling and predicting with Machine Learning

The main goal of the entire project is to predict heart disease occurrence with the highest accuracy. In order to achieve this, we will test several classification algorithms. This section includes all results obtained from the study and introduces the best performer according to accuracy metric. We have chosen several algorithms typical for solving supervised learning problems throughout classification methods.

First of all, let's equip ourselves with a handy tool that benefits from the cohesion of SciKit Learn library and formulate a general function for training our models. The reason for displaying accuracy on both, train and test sets, is to allow us to evaluate whether the model overfits or underfits the data (so-called bias/variance tradeoff).

```python
def train_model(X_train, y_train, X_test, y_test, classifier, **kwargs):

    """
    Fit the chosen model and print out the score.

    """

    # instantiate model
    model = classifier(**kwargs)

    # train model
    model.fit(X_train,y_train)

    # check accuracy and print out the results
    fit_accuracy = model.score(X_train, y_train)
    test_accuracy = model.score(X_test, y_test)

    print(f"Train accuracy: {fit_accuracy:0.2%}")
    print(f"Test accuracy: {test_accuracy:0.2%}")

    return model
```

### 3.8.1 Logistic Regression

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Unlike linear regression which outputs continuous number values, logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes.

**Types of logical regression:**

- Binary (Pass/Fail)
- Multi (Cats, Dogs, Sheep)

**Sigmoid function**

S(z)=1/1+e^−z

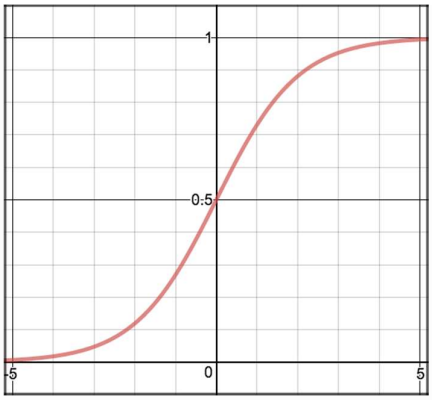



**Figure 29** sigmoid function

## Decision Boundary

p≥0.5, class=1 p<0.5, class=0

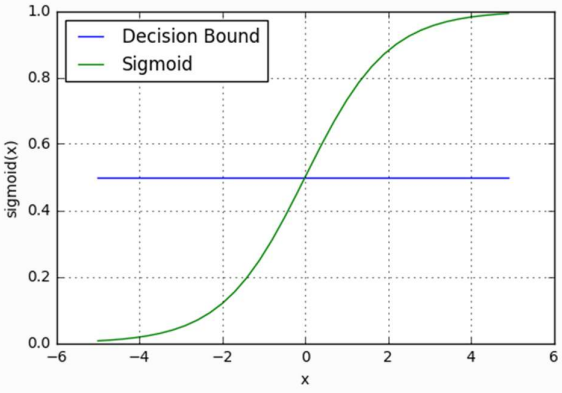



**Figure 30** Decision Boundary

## Cost Function

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \text{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

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

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

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

**Vectorized cost function**

$$h = g(X\theta)$$
$$J(\theta) = \frac{1}{m} \cdot \left(-y^T \log(h) - (1-y)^T \log(1-h)\right)$$

**For Multiclass -** Instead of y=0,1 we will expand our definition so that y=0,1...n. Basically we re-run binary classification multiple times, once for each class.

**Procedure -**

1. Divide the problem into n+1 binary classification problem (+1 because the index starts at 0).
2. For each class…
3. Predict the probability the observations are in that single class.
4. prediction = <math>max (probability of the classes)

```
# Logistic Regression
model = train_model(X_train, Y_train, X_test, Y_test, LogisticRegression)
```

```
Train accuracy: 84.71%
Test accuracy: 85.25%
```

**Accuracy score of Logistic Regression is: 85.25%**

**Confusion Matrix**

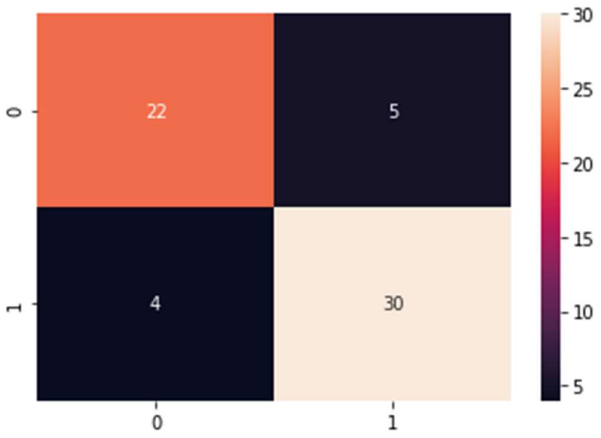



**Figure 31** logistic regression confusion matrix

### 3.8.1.2 precision score

```
[ ]  1 from sklearn.metrics import precision_score

[ ]  1 precision = precision_score(Y_test, Y_pred_lr)

[ ]  1 print("Precision: ",precision)

Precision:  0.8571428571428571
```

### Recall

```
[ ]  1 from sklearn.metrics import recall_score

[ ]  1 recall = recall_score(Y_test, Y_pred_lr)

[ ]  1 print("Recall is: ",recall)

Recall is:  0.8823529411764706
```

**F-Score**

balance of precision and recall score

```
[ ]  1 print((2*precision*recall)/(precision+recall))
```

```
0.8695652173913043
```

**false negative**

```
[ ]  1 #let us save TP, TN, FP, FN
     2 TN=CM.iloc[0,0]
     3 FP=CM.iloc[0,1]
     4 FN=CM.iloc[1,0]
     5 TP=CM.iloc[1,1]
```

- **false negative rate of the model**

```
[ ]  1 fnr=FN*100/(FN+TP)
     2 fnr
```

```
11.764705882352942
```

### 3.8.2 Random Forest

Random Forest is a supervised learning algorithm. Random forest can be used for both classification and regression problems, by using random forest regressor we can use random forest on regression problems. But we have used random forest on classification in this project so we will only consider the classification part.

**Random Forest pseudocode**

1. Randomly select **"k"** features from total **"m"** features.

   Where **k << m**

2. Among the **"k"** features, calculate the node **"d"** using the best split point.
3. Split the node into **daughter nodes** using the **best split**.
4. Repeat **1 to 3** steps until "l" number of nodes has been reached.

5. Build forest by repeating steps **1 to 4** for "n" number times to create **"n" number of trees**.

**Random forest prediction pseudocode**

1. Takes the **test features** and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target)
2. Calculate the **votes** for each predicted target.
3. Consider the **high voted** predicted target as the **final prediction** from the random forest algorithm.

```
#Random forest with 100 trees
from sklearn.ensemble import RandomForestClassifier
rf = RandomForestClassifier(n_estimators=100, random_state=0)
rf.fit(X_train, Y_train)
print("Accuracy on training set: {:.3f}".format(rf.score(X_train, Y_train)))
print("Accuracy on test set: {:.3f}".format(rf.score(X_test, Y_test)))
```

```
Accuracy on training set: 1.000
Accuracy on test set: 0.885
```

**Pruning the depth of the trees**

```
rf1 = RandomForestClassifier(max_depth=3, n_estimators=100, random_state=0)
rf1.fit(X_train, Y_train)
print("Accuracy on training set: {:.3f}".format(rf1.score(X_train, Y_train)))
print("Accuracy on test set: {:.3f}".format(rf1.score(X_test, Y_test)))
```

```
Accuracy on training set: 0.876
Accuracy on test set: 0.869
```

Accuracy score of Random Forest is 86.9%

## Confusion Matrix

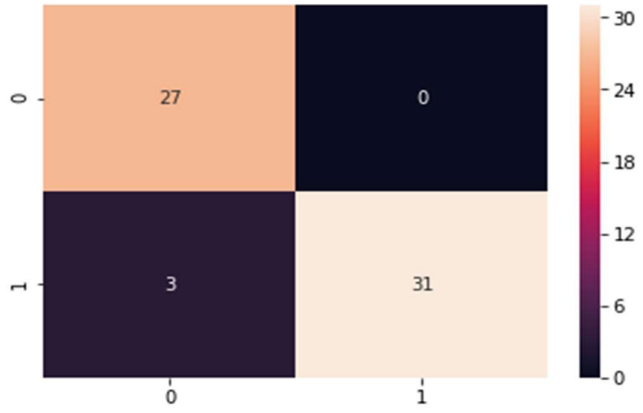



**Figure 32** Random forest confusion matrix

## Precision score

```
[ ]  1 from sklearn.metrics import precision_score

[ ]  1 precision = precision_score(Y_test, Y_pred_rf)

[ ]  1 print("Precision: ",precision)
```

```
Precision:  1.0
```

## Recall

```
[ ]  1 from sklearn.metrics import recall_score

[ ]  1 recall = recall_score(Y_test, Y_pred_rf)

[ ]  1 print("Recall is: ",recall)
```

```
Recall is:  0.9117647058823529
```

## F-Score

```
[ ]  1 print((2*precision*recall)/(precision+recall))
```

```
0.9538461538461539
```

**False negative rate**

```
[ ]  1 fnr=FN*100/(FN+TP)
     2 fnr
```

```
8.823529411764707
```

### 3.8.3 Naïve Bayes

**Bayes' Theorem** is stated as:

***P(h|d) = (P(d|h) * P(h)) / P(d)***

- **P(h|d)** is the probability of hypothesis h given the data d. This is called the posterior probability.
- **P(d|h)** is the probability of data d given that the hypothesis h was true.
- **P(h)** is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h.
- **P(d)** is the probability of the data (regardless of the hypothesis).

we are interested in calculating the posterior probability of P(h|d) from the prior probability p(h) with P(D) and P(d|h). After calculating the posterior probability for a number of different hypotheses, we will select the hypothesis with the highest probability. This is the maximum probable hypothesis and may formally be called the (MAP) hypothesis.
This can be written as:

***MAP(h) = max(P(h|d))***

***or***

***MAP(h) = max((P(d|h) * P(h)) / P(d))***

***or***

$$MAP(h) = max(P(d|h) * P(h))$$

The P(d) is a normalizing term which allows us to calculate the probability. We can drop it when we are interested in the most probable hypothesis as it is constant and only used to normalize. Back to classification, if we have an even number of instances in each class in our training data, then the probability of each class (e.g. P(h)) will be equal. Again, this would be a constant term in our equation, and we could drop it so that we end up with:

$$MAP(h) = max(P(d|h))$$

Naive Bayes is a classification algorithm for binary (two-class) and multi-class classification problems. The technique is easiest to understand when described using binary or categorical input values. It is called *naive Bayes* or *idiot Bayes* because the calculation of the probabilities for each hypothesis are simplified to make their calculation tractable. Rather than attempting to calculate the values of each attribute value P (d1, d2, d3|h), they are assumed to be conditionally independent given the target value and calculated as P(d1|h) * P(d2|H) and so on. This is a very strong assumption that is most unlikely in real data, i.e. that the attributes do not interact. Nevertheless, the approach performs surprisingly well on data where this assumption does not hold.

$$MAP(h) = max(P(d|h) * P(h))$$

**Gaussian Naïve Bayes**

$$mean(x) = 1/n * sum(x)$$

Where n is the number of instances and x are the values for an input variable in your training data. We can calculate the standard deviation using the following equation:

$$standard\ deviation(x) = sqrt\ (1/n * sum(xi-mean(x)^2))$$

This is the square root of the average squared difference of each value of x from the mean value of x, where n is the number of instances, sqrt() is the square root function, sum() is the sum function, xi is a specific value of the x variable for the i'th instance and mean(x) is described above, and ^2 is the square. Gaussian PDF with a new input for the variable, and in

return the Gaussian PDF will provide an estimate of the probability of that new input value for that class.

***pdf (x, mean, sd) = (1 / (sqrt (2 * PI) * sd)) * exp (-((x-mean^2)/(2*sd^2)))***

Where pdf(x) is the Gaussian Probability Density Function (PDF), sqrt () is the square root, mean and sd are the mean and standard deviation calculated above, Pi is the numerical constant, exp () is the numerical constant e or Euler's number raised to power and x is the input value for the input variable.

```
#Gaussian Naive Bayes
model = train_model(X_train, Y_train, X_test, Y_test, GaussianNB)
```

```
Train accuracy: 83.47%
Test accuracy: 85.25%
```

**Confusion matrix**

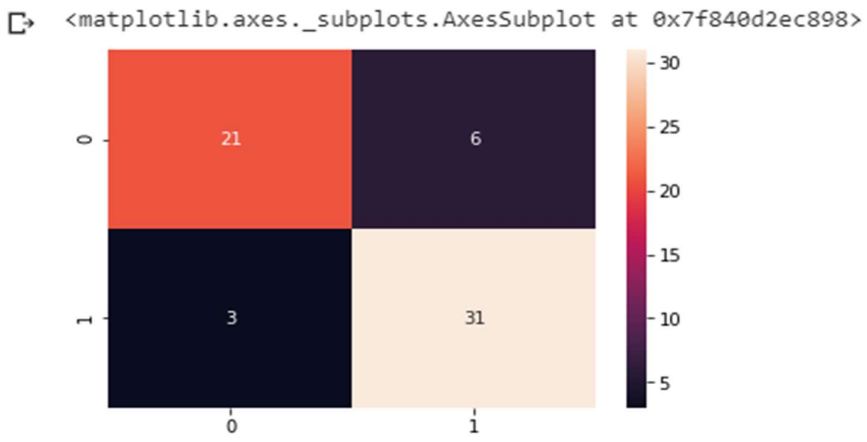



**Figure 33** Naive Bayes confusion matrix

**Precision score**

```
print("Precision: ",precision)
```

```
Precision:  0.8378378378378378
```

**Recall**

```
[ ]  1 print("Recall is: ",recall)
```

```
Recall is:  0.9117647058823529
```

**F-Score**

```
0.8732394366197184
```

**False negative**

```
[ ]  1 fnr = FN*100/(FN+TP)
     2 fnr
```

```
8.823529411764707
```

### 3.8.4 K-Nearest Neighbor

We can implement a KNN model by following the below steps:

1. Load the data
2. Initialize the value of k
3. For getting the predicted class, iterate from 1 to total number of training data points

    - Calculate the distance between test data and each row of training data. Here we will use Euclidean distance as our distance metric since it's the most popular method. The other metrics that can be used are Chebyshev, cosine, etc.
    - Sort the calculated distances in ascending order based on distance values
    - Get top k rows from the sorted array
    - Get the most frequent class of these rows
    - Return the predicted class

```
# KNN
model = train_model(X_train, Y_train, X_test, Y_test, KNeighborsClassifier)
```

```
Train accuracy: 78.10%
Test accuracy: 63.93%
```

**Using different inputs –**

It turns out that the value of n = 8 is optimal

```
# Seek optimal 'n_neighbours' parameter
for i in range(1,10):
    print("n_neigbors = "+str(i))
    train_model(X_train, Y_train, X_test, Y_test, KNeighborsClassifier, n_neighbors=i)
```

```
n_neigbors = 1
Train accuracy: 100.00%
Test accuracy: 52.46%
n_neigbors = 2
Train accuracy: 79.75%
Test accuracy: 59.02%
n_neigbors = 3
Train accuracy: 78.10%
Test accuracy: 63.93%
n_neigbors = 4
Train accuracy: 76.03%
Test accuracy: 63.93%
n_neigbors = 5
Train accuracy: 78.10%
Test accuracy: 63.93%
n_neigbors = 6
Train accuracy: 74.38%
Test accuracy: 65.57%
n_neigbors = 7
Train accuracy: 72.31%
Test accuracy: 67.21%
n_neigbors = 8
Train accuracy: 71.90%
Test accuracy: 68.85%
n_neigbors = 9
Train accuracy: 73.14%
Test accuracy: 67.21%
```

### Confusion matrix

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f840d207358>
```

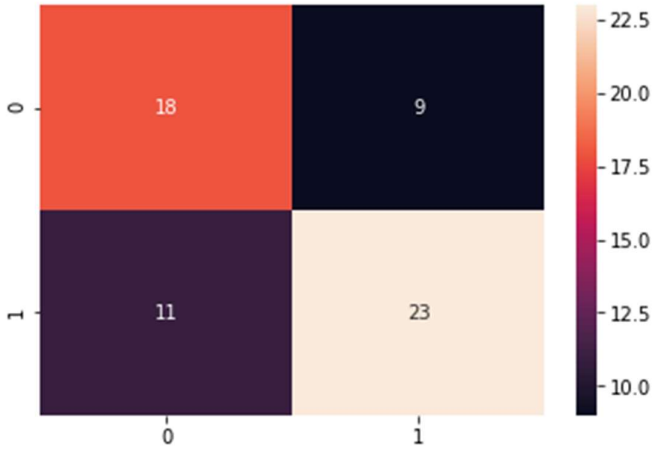



**Figure 34** KNN confusion matrix

### Precision score

```
[ ]  1 print("Precision: ",precision)
```

```
Precision:  0.71875
```

### Recall

```
[ ]  1 print("Recall is: ",recall)
```

```
Recall is:  0.6764705882352942
```

### F-Score

```
0.696969696969697
```

### False negative

```
[ ]  1 fnr = FN*100/(FN+TP)
     2 fnr
```

```
32.35294117647059
```

### 3.8.5 Decision Tree

**Pseudocode**

1. Place the best attribute of the dataset at the **root** of the tree.
2. Split the training set into **subsets**. Subsets should be made in such a way that each subset contains data with the same value for an attribute.
3. Repeat step 1 and step 2 on each subset until you find **leaf nodes** in all the branches of the tree.

**Assumptions while creating Decision Tree**

- At the beginning, the whole training set is considered as the **root.**
- Feature values are preferred to be categorical. If the values are continuous then they are discretized prior to building the model.
- Records are **distributed recursively** on the basis of attribute values.
- Order to placing attributes as root or internal node of the tree is done by using some statistical approach.

**The popular attribute selection measures**

- Information gain
- Gini index

**Attribute selection method -** A dataset consists of **"n"** attributes then deciding which attribute to place at the root or at different levels of the tree as internal nodes is a complicated step. By just randomly selecting any node to be the root can't solve the issue. If we follow a random approach, it may give us bad results with low accuracy. For solving this attribute selection problem, researchers worked and devised some solutions. They suggested using some *criterion* like **information gain, Gini index,** etc. These criterions will calculate values for every attribute. The values are sorted, and attributes are placed in the tree by following the order i.e., the attribute with a high value (in case of information gain) is placed at the root. While using information Gain as a criterion, we assume attributes to be categorical, and for Gini index, attributes are assumed to be continuous.[132]

**Gini Index -** Gini Index is a metric to measure how often a randomly chosen element would be incorrectly identified. It means an attribute with lower Gini index should be preferred.

$$Gini\ Index = 1 - \sum_j p_j^2$$

```
from sklearn.tree import DecisionTreeClassifier
tree1 = DecisionTreeClassifier(random_state=0)
tree1.fit(X_train, Y_train)
print("Accuracy on training set: {:.3f}".format(tree1.score(X_train, Y_train)))
print("Accuracy on test set: {:.3f}".format(tree1.score(X_test, Y_test)))
```

```
Accuracy on training set: 1.000
Accuracy on test set: 0.787
```

The accuracy on the training set is 100%, while the test set accuracy is much worse. This is an indicative that the tree is overfitting and not generalizing well to new data. Therefore, we need to apply pre-pruning to the tree. We set max depth=3, limiting the depth of the tree decreases overfitting. This leads to a lower accuracy on the training set, but an improvement on the test set.

```
tree1 = DecisionTreeClassifier(max_depth=3, random_state=0)
tree1.fit(X_train, Y_train)
print("Accuracy on training set: {:.3f}".format(tree1.score(X_train, Y_train)))
print("Accuracy on test set: {:.3f}".format(tree1.score(X_test, Y_test)))
```

```
Accuracy on training set: 0.843
Accuracy on test set: 0.820
```

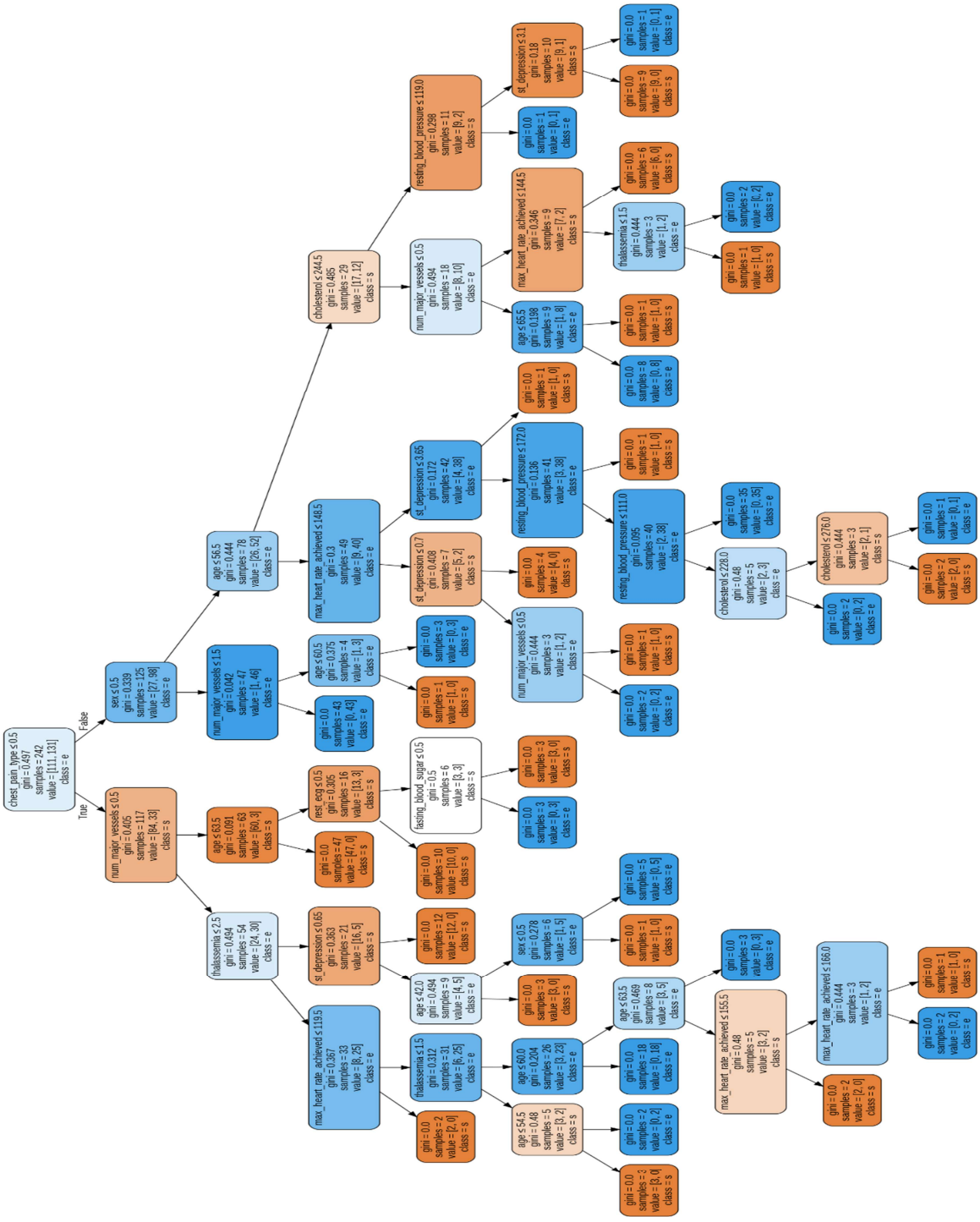



**Figure 35** Decision tree visualization

### Confusion matrix

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f840d18c5f8>
```

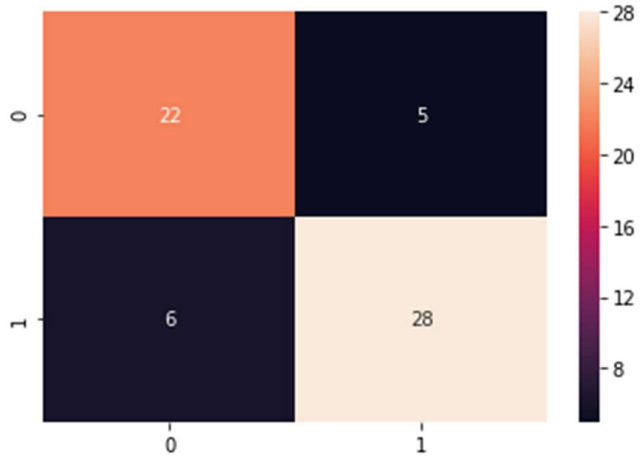



**Figure 36** decision tree confusion matrix

### Precision score

```
[ ]  1 print("Precision: ",precision)
```

```
Precision:  0.8484848484848485
```

### Recall

```
[ ]  1 print("Recall is: ",recall)
```

```
Recall is:  0.8235294117647058
```

### F-Score

```
0.8358208955223881
```

### False negative rate

```
17.647058823529413
```

## 3.9 Result

```
# initialize an empty list
accuracy = []

# list of algorithms names
classifiers = ['KNN', 'Decision Trees', 'Logistic Regression', 'Naive Bayes', 'Random Forests']

# list of algorithms with parameters
models = [KNeighborsClassifier(n_neighbors=8), DecisionTreeClassifier(max_depth=3, random_state=0), Logis
        GaussianNB(), RandomForestClassifier(max_depth=3, n_estimators=100, random_state=0)]

# loop through algorithms and append the score into the list
for i in models:
    model = i
    model.fit(X_train, Y_train)
    score = model.score(X_test, Y_test)
    accuracy.append(score)
```

```
# create a dataframe from accuracy results
summary = pd.DataFrame({'accuracy':accuracy}, index=classifiers)
summary
```

| | accuracy |
|---|---|
| **KNN** | 0.688525 |
| **Decision Trees** | 0.819672 |
| **Logistic Regression** | 0.852459 |
| **Naive Bayes** | 0.852459 |
| **Random Forests** | 0.868852 |

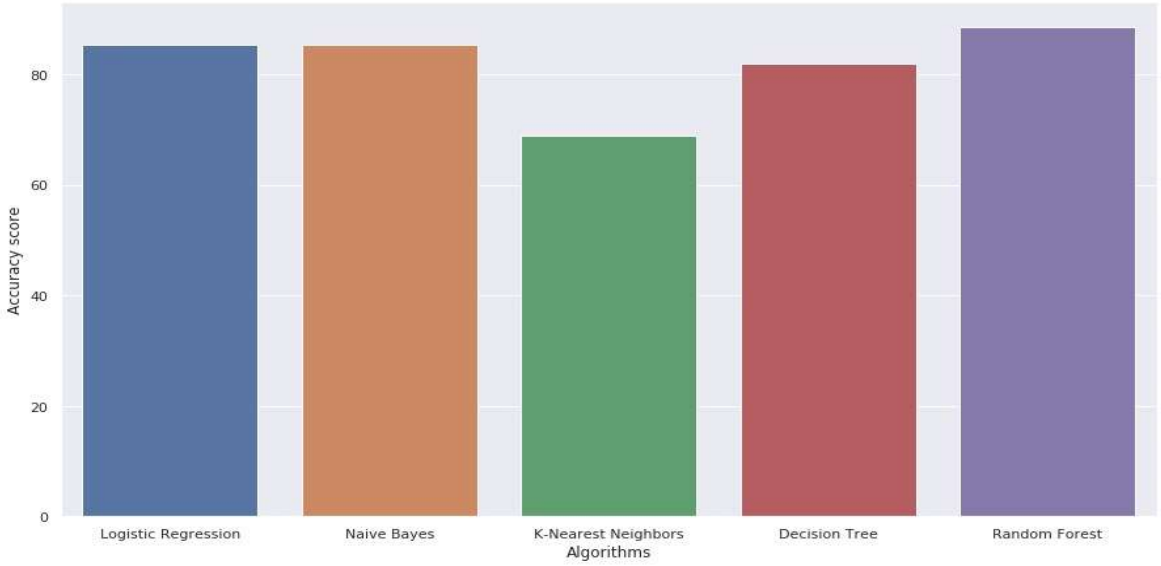



**Figure 37** Accuracy Scores of all Algorithms

## 3.10 System Requirements

1. Intel® Core™ i5 processor 8250U at 1.60 GHz or 1.80 GHz, 8 GB of DRAM.
2. Disk space 2TB.
3. Operating System: 64-bit Windows 10 Pro.

**Recommended System Requirements**

- Processors: Intel® Core™ i5 processor 4300M at 2.60 GHz or 2.59 GHz (1 socket, 2 cores, 2 threads per core), 8 GB of DRAM.
- Operating systems: Windows® 10, macOS*, and Linux*

**Minimum System Requirements**

- Processors: Intel Atom® processor or Intel® Core™ i3 processor
- Disk space: 1 GB
- Operating systems: Windows* 7 or later, macOS, and Linux
- Python* versions: 2.7.X, 3.6.X
- Included development tools: conda*, conda-env, Jupyter Notebook* (IPython)
- Compatible tools: Microsoft Visual Studio*, PyCharm*Included Python packages - NumPy, SciPy, scikit-learn*, pandas, Matplotlib, Numba*, Intel® Threading Building Blocks, pyDAAL, Jupyter, mpi4py, PIP*, and others.

**Software**

- PIP and NumPy: Ubuntu*, Python 3.6.2, NumPy 1.13.1, scikit-learn 0.18.2
- Windows: Python 3.6.2, PIP and NumPy 1.13.1, scikit-learn 0.18.2
- Intel® Distribution for Python* 2018
-

**Modifications**

- Scikit-learn: Conda*-installed NumPy with Intel® Math Kernel Library (Intel® MKL) on Windows (PIP-installed SciPy on Windows contains Intel MKL dependency).

# CHAPTER: 4 CONCLUSION

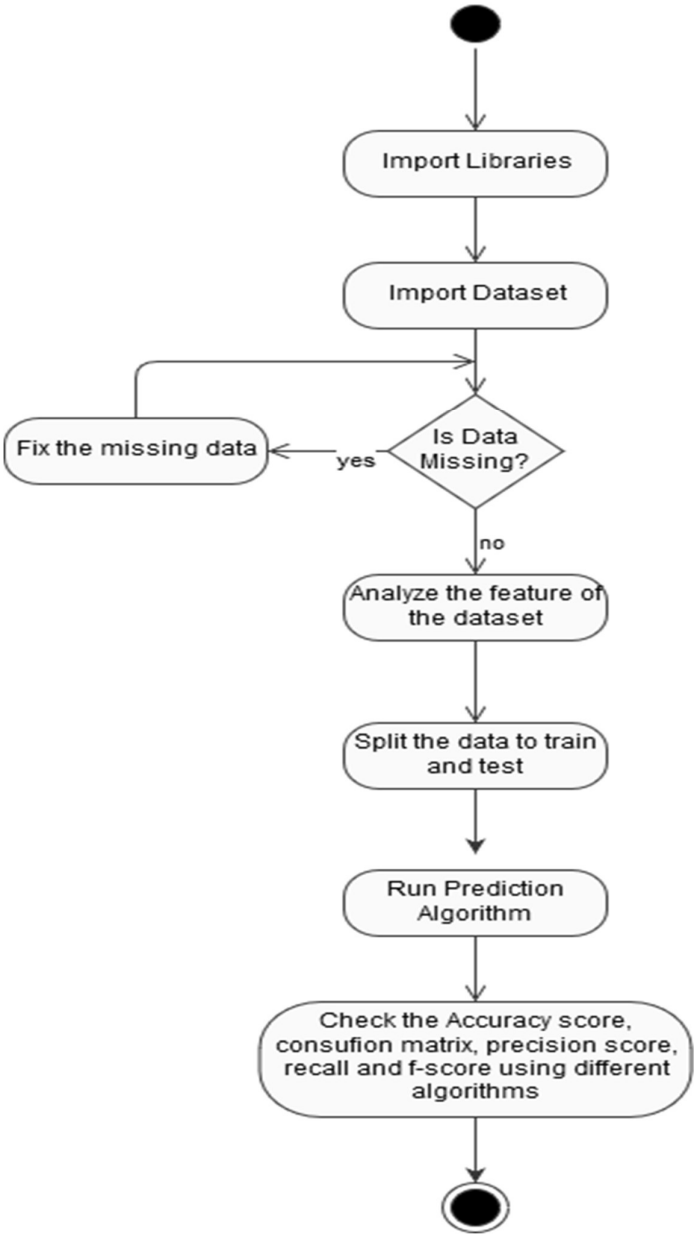
## 4.1 Inference

The overall objective of our project is to predict accurately with less number of tests and attributes the presence of heart disease. In this project, fourteen attributes are considered which form the primary basis for tests and give accurate results more or less. Many more input attributes can be taken but our goal is to predict with less number of attributes and faster efficiency to predict the risk of having heart disease at a particular age span. Five data mining classification techniques were applied namely K-Nearest Neighbor, Naive Bayes, Decision Tree, Random Forest & Logistic Regression. It is shown that Random Forest has better accuracy than the other techniques.

This is the most effective model to predict patients with heart disease. This project could answer complex queries, each with its own strength with respect to ease of model interpretation, access to detailed information and accuracy.

This project can be further enhanced and expanded. For example, it can incorporate other medical attributes besides the 14 attributes we used. It can also incorporate other data mining techniques, e.g., Time Series, Clustering and Association Rules. Continuous data can also be used instead of just categorical data. Another area is to use Text Mining to mine the vast amount of unstructured data available.

This project is presented using data mining techniques. From logistic regression, KNN, Naive Bayes, Decision Tree, Random forest are used to develop the system. Random Forest proves the better results and assists the domain experts and even the person related to the medical field to plan for a better and early diagnosis for the patient. This system performs realistically well even without retraining




**ACTIVITY DIAGRAM**

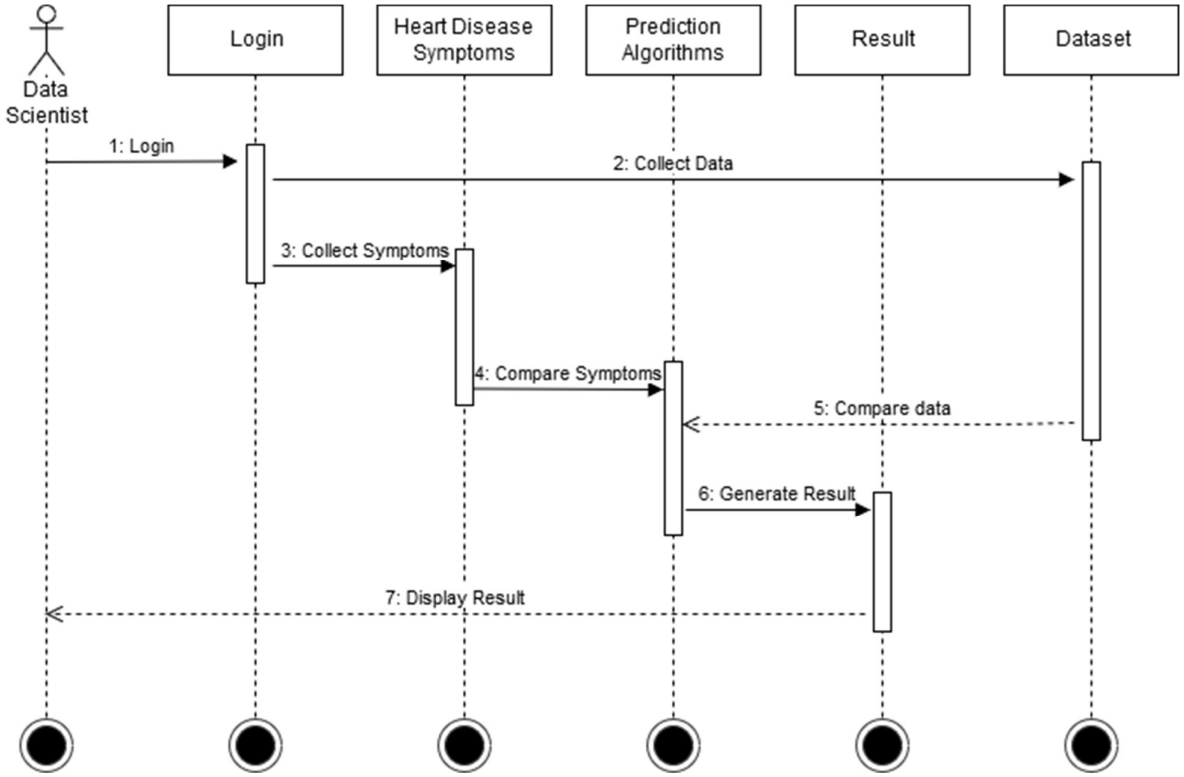
## 4.2 Drawbacks

The Algorithms used in our project does not give a 100% accuracy, so the prediction is not 100% feasible. Clinical diagnosis and diagnosis using our project may differ slightly because the prediction is not 100% accurate. Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. The automation of the same would be highly beneficial. Clinical decisions are often made based on doctor's intuition and experience rather than on the knowledge rich data collected from the dataset.

## 4.3 Future Scope

We are planning to introduce an efficient disease prediction system to predict the heart disease with better accuracy using Support Vector Machine (SVM). Our project aims to provide a web platform to predict the occurrences of disease based on various symptoms. The user can select various symptoms and can find the diseases with their probabilistic figures. Our project can be improved by implementing medicine suggestion to the patient along with the results. We can implement a feedback from the experienced doctors who can give their views and opinions about certain medicines /practices done by the doctor on the patient. We can implement a live chat option where the patient can chat with a doctor available regarding medication for the respective result for their symptoms. Our project could be used as a training tool for Nurses and Doctors who are freshly introduced in the field related to heart diseases. The patient can have a choice in choosing the medicines he/she should take in order to have a healthier life. Moreover, if implemented on a large scale it can be used in medical facilities like hospital, clinics where a patient wouldn't have to wait in long queues for treatment if he is feeling symptoms related to heart disease.

## Sequence Diagram for Future Scope



# REFERENCES


[1] M. K. Awang and F. Siraj, "Utilization of an artificial neural network in the prediction of heart disease," *Int. J. Bio-Science Bio-Technology*, vol. 5, no. 4, pp. 159–165, 2013.

[2] P. Selvakumar and S. P. Rajagopalan, "A survey on neural network models for heart disease prediction," *J. Theor. Appl. Inf. Technol.*, vol. 67, no. 2, pp. 485–497, 2014.

[3] A. Methaila, P. Kansal, H. Arya, and P. Kumar, "Early Heart Disease Prediction Using Data Mining Techniques," vol. 6956, no. October, pp. 53–59, 2014.

[4] I. S. F. Dessai, "Intelligent Heart Disease Prediction System Using Probabilistic Neural Network," no. 5, pp. 38–44, 2013.

[5] T. Karayilan and Ö. Kiliç, "Prediction of Heart disease using neural network," in *2nd International Conference on Computer Science and Engineering, UBMK 2017*, 2017, pp. 719–723.

[6] M. Mardiyono, R. Suryanita, and A. Adnan, "Intelligent Monitoring System on Prediction of Building Damage Index using Neural-Network," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 10, no. 1, p. 155, 2015.

[7] N. Guru, A. Dahiya, and N. Rajpal, "Decision support system for heart disease diagnosis using Neural Network," 2007.

[8] J. S. Singh Navdeep, "No Title," *Intell. Hear. Dis. Predict. Syst. using CANFIS Genet. Algorithm, Int. J. Biol. Med. Sci.*, no. International Journal of Advance Research, Ideas and Innovations in Technology© 2018, www.IJARIIT.comAll Rights, p. Reserved Page | 987, 2008.

[9] R. G. S. Rajkumar Asha, "No Title," *Diagnosis Hear. Dis. Using Datamining Algorithm*, vol. Issue 10 V, no. Global Journal of Computer Science and Technology, p. Page 38, 2010.

[10] S. Radhimeenakshi and G. M. Nasira, "Prediction of Heart Disease using Neural Network with Back Propagation," *Int. J. Data Min. Tech. Appl.*, vol. 4, no. 1, pp. 19–22, 2016.

[11] "Prediction of Heart Disease using Artificial Neural Network," *VFAST Trans. Softw. Eng.*, pp. 102–112, 2019.

[12] Z. M. A. Mehta Rupa G. ,Rana Dipti P., "No Title," *A Nov. Fuzzy Based Classif. Data Min. using Fuzzy Discret.*, no. World Congress on Computer Science and Information, 2009.

[13] A. A. Shinde, S. N. Kale, R. M. Samant, A. S. Naik, and S. A. Ghorpade, "Heart Disease Prediction System using Multilayered Feed Forward Neural Network and Back Propagation Neural Network," *Int. J. Comput. Appl.*, vol. 166, no. 7, pp. 975–8887, 2017.

[14] N. Al-Milli, "Backpropogation neural network for prediction of heart disease," *J. Theor. Appl. Inf. Technol.*, vol. 56, no. 1, pp. 131–135, 2013.

[15] G. S. Kumari Milan, "No Title," *Comp. Study Data Min. Classif. Methods Cardiovasc. Dis. Predict.*, vol. Vol. 2, no. IJCST Vol. 2, Iss ue2, June 2011, 2011.

[16] C. Ordonez, "Improving Heart Disease Prediction using Constrained Association Rules," *Tech. Semin. Present. Univ. Tokyo*, 2004.

[17] M. C. and P. M. Franck Le Duff, CristianMunteanb, "Predicting Survival Causes After Out of Hospital Cardiac Arrest using Data Mining Method," *Stud. Health Technol. Inform.*, vol. Vol. 107, no. 2, p. No. 2, pp. 1256–1259, 2004.

[18] W. J. F. and G. Piatetsky-Shapiro, "Knowledge Discovery in Databases: An Overview," *AI Mag.*, vol. Vol. 13, N, no. 3, pp. 57–70, 1996.

[19] K. Y. N. and K. H. R. Heon Gyu Lee, "Mining Bio Signal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV," *Proc. Int. Conf. Emerg. Technol. Knowl. Discov. Data Min.*, p. pp. 56–66, 2007.

[20] B. J. L. and K. H. R. Kiyong Noh, HeonGyu Lee, Ho-Sun Shon, "Associative Classification Approach for Diagnosing Cardiovascular Disease," *Intell. Comput. Signal Process. Pattern Recognit.*, vol. 345, pp. 721–727, 2006.

[21] L. P. and R. Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm," *Int. J. Biol. Biomed. Med. Sci.*, vol. Vol. 3, no. No. 3, pp. 1-8, 2008.

[22] A. D. and N. R. Niti Guru, "Decision Support System for Heart Disease Diagnosis using Neural Network ," *Delhi Bus. Rev.*, vol. Vol. 8, no. 1, pp. 1–6.

[23] S. P. and R. Awang, "Intelligent Heart Disease Prediction System using Data Mining

Techniques," *Int. J. Comput. Sci. Netw. Secur.*, vol. Vol. 8, no. No. 8, p. pp. 1–6, 2008.

[24] Shantakumar B. Patil and Y.S. Kumaraswamy, "'Intelligent and Effective Heart Attack Prediction System using Data Mining and Artificial Neural Network'," *Eur. J. Sci. Res.*, vol. Vol. 31, no. No. 4, p. pp. 642-656, 2009.

[25] X. Y. et Al., "Combination Data Mining Models with New Medical Data to Predict Outcome of Coronary Heart Disease," *Proc. Int. Conf. Converg. Inf. Technol.*, pp. 868–872, 2007.

[26] E. Y. and C. Kilikcier, "Determination of Patient State from Cardiotocogram using LS-SVM with Particle Swarm Optimization and Binary Decision Tree," *Master Thesis, Dep. Electr. Electron. Eng.*, no. Uludag, 2013.

[27] N. S. and D. Singh, "Performance Evaluation of K-Means and Hierarchal Clustering in Terms of Accuracy and Running Time," *Dep. Comput. Sci. Eng. Barkatullah Univ. Inst. Technol.*, no. Ph.D Dissertation, 2012.

[28] S. G. Manoj B, Kumar G, Ramesh G, "Emerging risk factors for cardiovascular diseases: Indian context.," *Indian J Endocrinol Metab*, vol. 17, pp. 806–814, 2013.

[29] Elama Zannatul F., "Combination of Naive Bayes classifier and K-NN in the classification based classification models.," *Comput. Inf. Sci*, no. 6, pp. 48–56, 2013.

[30] Halaudi Daniel M., "Prediction of heart disease using classification algorithms.," *WCSECS*, pp. 22–24, 2014.

[31] A. Uma ND, "Extraction of action rules for chronic kidney disease using Naive Bayes classifier.," *IEEE Int Conf. Comput Intell. Comput Res*, 2016.

[32] J. A. Marx *et al.*, *Rosen's emergency medicine : concepts and clinical practice*, Eighth. Philadelphia, PA: Elsevier/Saunders, 2014.

[33] M. S., "Integrating Naive and clustering with a different initial centroid selection methods in the diagnosis of heart disease prediction.," *CS IT CSCP*, pp. 125–137, 2012.

[34] Rupali RP MS., "Heart disease prediction system using naive based and Jelmeck Mercer smoothing.," *IJARCCE*, no. 3, pp. 6787–6792, 2014.

[35] C. Aflori, M. Craus, A.J.T. Lee, Y.H. Liu, H.Mu Tsai, H.-Hui Lin, H-W. Wu, "Grid implementation of the Apriori algorithm Advances in Engineering Software," *Min.*

*Freq. patterns image databases with 9D-SPA Represent. - J. Syst. Softw.*, vol. Volume 38, no. Issue 5, pp. 295-300.

[36] K. Srinivas, "Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques," *IEEE Trans. Comput. Sci. Educ. (ICCSE),* p. p(1344-1349), 2010.

[37] Y. S. K. Shanta kumar, B.Patil, "Predictive data mining for medical diagnosis of heart disease prediction," *IJCSE*, vol. Vol .17, 2011.

[38] M. Anbarasi et. al., "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm," *Int. J. Eng. Sci. Technol.*, vol. 2, no. 10, pp. 5370–5376, 2010.

[39] H. W. Khaing, "Data Mining based Fragmentation and Prediction of Medical Data," *IEEE,* 2011.

[40] V. Chauraisa and S. Pal, "Data Mining Approach to Detect Heart Diseases," *Int. J. Adv. Comput. Sci. Inf. Technol.*, vol. Vol. 2, no. No. 4, p. pp 56-66., 2013.

[41] Q. J., "Induction of decision trees.," *Mach Learn*, vol. 1, p. 81—106, 1986.

[42] V. L. Roger *et al.*, "Heart disease and stroke statistics-2011 update: A report from the American Heart Association," *Circulation*, 2011.

[43] R. S. Vasan *et al.*, "Relative importance of borderline and elevated levels of coronary heart disease risk factors," *Annals of Internal Medicine*. 2005.

[44] R. B. D'Agostino *et al.*, "General cardiovascular risk profile for use in primary care: The Framingham heart study," *Circulation*, 2008.

[45] NCEP - National Cholesterol Education Program, "Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on," *01-3670*, 2001.

[46] S. M. Grundy *et al.*, "Implications of recent clinical trials for the National Cholesterol Education Program Adult Treatment Panel III Guidelines.," in *Journal of the American College of Cardiology*, 2004.

[47] J. Hippisley-Cox, C. Coupland, Y. Vinogradova, J. Robson, M. May, and P. Brindle, "Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: Prospective open cohort study," *Br. Med. J.*, 2007.

[48] M. Woodward, P. Brindle, and H. Tunsfall-Pedoe, "Adding social deprivation and family history to cardiovascular risk assessment: The ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC)," *Heart*, 2007.

[49] R. Jackson, "Updated New Zealand cardiovascular disease risk-benefit prediction guide," *BMJ*, 2000.

[50] X. F. Zhang, J. Attia, C. D'Este, X. H. Yu, and X. G. Wu, "A risk score predicted coronary heart disease and stroke in a Chinese cohort," *J. Clin. Epidemiol.*, 2005.

[51] P. M. Ridker, J. E. Buring, N. Rifai, and N. R. Cook, "Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: The Reynolds Risk Score," *J. Am. Med. Assoc.*, 2007.

[52] R. M. Conroy *et al.*, "Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project.," *Eur. Heart J.*, 2003.

[53] I. Graham *et al.*, "European guidelines on cardiovascular disease prevention in clinical practice: executive summary," *Eur. J. Cardiovasc. Prev. Rehabil.*, 2007.

[54] A. D. Sniderman and C. D. Furberg, "Age as a modifiable risk factor for cardiovascular disease," *Lancet*, 2008.

[55] D. M. Lloyd-Jones *et al.*, "Prediction of lifetime risk for cardiovascular disease by risk factor burden at 50 years of age," *Circulation*, 2006.

[56] D. F. Terry *et al.*, "Cardiovascular risk factors predictive for survival and morbidity-free survival in the oldest-old Framingham Heart Study participants," *J. Am. Geriatr. Soc.*, 2005.

[57] R. J. Goldberg, M. Larson, and D. Levy, "Factors associated with survival to 75 years of age in middle-aged men and women: The Framingham study," *Arch. Intern. Med.*, 1996.

[58] W. B.J. *et al.*, "Midlife risk factors and healthy survival in men," *J. Am. Med. Assoc.*, 2006.

[59] P. S. Yusuf *et al.*, "Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): Case-control study," *Lancet*, 2004.

[60] R. B. D'Agostino *et al.*, "General Cardiovascular Risk Profile for Use in Primary

Care," *Circulation*, 2008.

[61] R. S. Vasan and R. B. D'Agostino, "Age and Time Need Not and Should Not Be Eliminated From the Coronary Risk Prediction Models," *Circulation*, 2005.

[62] R. M. Epstein, B. S. Alper, and T. E. Quill, "Communicating evidence for participatory decision making," *Journal of the American Medical Association*. 2004.

[63] D. M. Lloyd-Jones *et al.*, "Lifetime risk for developing congestive heart failure: The Framingham Heart Study," *Circulation*, 2002.

[64] D. M. Lloyd-Jones, M. G. Larson, A. Beiser, and D. Levy, "Lifetime risk of developing coronary heart disease," *Lancet*, 1999.

[65] W. B. Kannel and R. S. Vasan, "Is Age Really a Non-Modifiable Cardiovascular Risk Factor?," *Am. J. Cardiol.*, 2009.

[66] S. Seshadri *et al.*, "The lifetime risk of stroke: Estimates from the framingham study," *Stroke*, 2006.

[67] J. Canto *et al.*, "Prevalence, clinical characteristics, and mortality among patients with myocardial infarction presenting without chest pain," *JAMA*, vol. 283, no. 24, pp. 3223–9, 2000.

[68] V. L. Roger *et al.*, "Executive summary: Heart disease and stroke statistics-2012 update: A report from the American heart association," *Circulation*, 2012.

[69] L. Mosca, H. Mochari-Greenberger, R. J. Dolor, L. K. Newby, and K. J. Robb, "Twelve-year follow-up of American women's awareness of cardiovascular disease risk and barriers to heart health," *Circ. Cardiovasc. Qual. Outcomes*, 2010.

[70] R. B. Jaffe, "Explaining the decrease in U.S. deaths from coronary disease, 1980-2000: Commentary," *Obstetrical and Gynecological Survey*. 2007.

[71] J. S. Berger *et al.*, "Sex differences in mortality following acute coronary syndromes," *JAMA - J. Am. Med. Assoc.*, 2009.

[72] K. P. Alexander *et al.*, "Sex differences in major bleeding with glycoprotein IIb/IIIa inhibitors: Results from the CRUSADE (Can rapid risk stratification of unstable angina patients suppress adverse outcomes with early implementation of the ACC/AHA guidelines) initiative," *Circulation*, 2006.

[73] S. O.M. *et al.*, "Improved clinical outcomes in patients undergoing coronary artery bypass grafting with coronary endarterectomy," *Ann. Thorac. Surg.*, 1999.

[74] J. D. Puskas, P. D. Kilgo, M. Kutner, S. V Pusca, O. Lattouf, and R. A. Guyton, "Off-pump techniques disproportionately benefit women and narrow the gender disparity in outcomes after coronary artery bypass surgery," *Circulation*, 2007.

[75] R. P. Hertz, A. N. Unger, J. A. Cornell, and E. Saunders, "Racial disparities in hypertension prevalence, awareness, and management," *Arch. Intern. Med.*, 2005.

[76] E. L. Barrett Connor, B. A. Cohn, D. L. Wingard, and S. L. Edelstein, "Why Is Diabetes Mellitus a Stronger Risk Factor for Fatal Ischemic Heart Disease in Women Than in Men?: The Rancho Bernardo Study," *JAMA J. Am. Med. Assoc.*, 1991.

[77] C. C. Cowie *et al.*, "Full accounting of diabetes and pre-diabetes in the U.S. population in 1988-1994 and 2005-2006," *Diabetes Care*, 2009.

[78] R. B. Ervin, "Prevalence of metabolic syndrome among adults 20 years of age and over, by sex, age, race and ethnicity, and body mass index: United States, 2003-2006.," *Natl. Health Stat. Report.*, 2009.

[79] K. M. Flegal, M. D. Carroll, C. L. Ogden, and L. R. Curtin, "Prevalence and trends in obesity among US adults, 1999-2008," *JAMA - J. Am. Med. Assoc.*, 2010.

[80] J. R. Pleis, J. W. Lucas, and W. BW, "Summary health statistics for U.S. adults: National Health Interview Survey, 2008.," *Vital Health Stat. 10.*, 2009.

[81] L. L. Leape, J. S. Weissman, E. C. Schneider, R. N. Piana, C. Gatsonis, and A. M. Epstein, "Adherence to practice guidelines: The role of specialty society guidelines," *Am. Heart J.*, 2003.

[82] T. M. Shaneyfelt, M. F. Mayo-Smith, and J. Rothwangl, "Are guidelines following guidelines?. The methodological quality of clinical practice guidelines in the peer-reviewed medical literature," *J. Am. Med. Assoc.*, 1999.

[83] M. D. Cabana and C. Kim, "Physician adherence to preventive cardiology guidelines for women," in *Women's Health Issues*, 2003.

[84] L. Mosca *et al.*, "National study of physician awareness and adherence to cardiovascular disease prevention guidelines," *Circulation*. 2005.

[85] J. Barnhart, V. Lewis, J. L. Houghton, and P. Charney, "Physician Knowledge Levels

and Barriers to Coronary Risk Prevention in Women. Survey Results from the Women and Heart Disease Physician Education Initiative," *Women's Heal. Issues*, 2007.

[86] W. R. Lewis *et al.*, "Trends in the use of evidence-based treatments for coronary artery disease among women and the elderly: Findings from the get with the guidelines quality-improvement program," *Circ. Cardiovasc. Qual. Outcomes*, 2009.

[87] A. H. Christian, T. Mills, S. L. Simpson, and L. Mosca, "Quality of cardiovascular disease preventive care and physician/practice characteristics," *J. Gen. Intern. Med.*, 2006.

[88] C. L. Bryson, R. R. Miller, A. E. Sales, B. Kopjar, and S. D. Fihn, "Adherence to Heart-Healthy Behaviors in a Sample of the U.S. Population," *Prev. Chronic Dis. Electron. Resour.*, 2005.

[89] T. G. Pickering *et al.*, "Recommendations for blood pressure measurement in humans and experimental animals. Part 1: Blood pressure measurement in humans: A statement for professionals from the subcommittee of professional and public education of the American Heart Association cou," *Hypertension*. 2005.

[90] W. B. Kannel, T. Gordon, and M. J. Schwartz, "Systolic versus diastolic blood pressure and risk of coronary heart disease. The Framingham study," *Am. J. Cardiol.*, 1971.

[91] J. Stamler and R. Stamler, "Intervention for the prevention and control of hypertension and atherosclerotic diseases: United States and international experience," *Am. J. Med.*, 1984.

[92] S. S. Franklin, S. A. Khan, N. D. Wong, M. G. Larson, and D. Levy, "Is pulse pressure useful in predicting risk for coronary heart disease? The Framingham Heart Study," *Circulation*, 1999.

[93] J. L. Izzo, D. Levy, and H. R. Black, "Importance of systolic blood pressure in older Americans," *Hypertension*, 2000.

[94] J. L. Probstfield, "Prevention of stroke by antihypertensive drug treatment in older persons with isolated systolic hypertension. Final results of the Systolic Hypertension in the Elderly Program (SHEP)," *Annals of Internal Medicine*. 1991.

[95] N. Sarwar *et al.*, "Diabetes mellitus, fasting blood glucose concentration, and risk of

vascular disease: A collaborative meta-analysis of 102 prospective studies," *Lancet*, 2010.

[96] R. F. Gillum and C. T. Grant, "Coronary heart disease in black populations II. Risk factors," *Am. Heart J.*, 1982.

[97] D. Steinberg, "Thematic review series: the pathogenesis of atherosclerosis: an interpretive history of the cholesterol controversy, part III: mechanistically defining the role of hyperlipidemia.," *J. Lipid Res.*, 2005.

[98] E. L. M. Barr *et al.*, "Risk of cardiovascular and all-cause mortality in individuals with diabetes mellitus, impaired fasting glucose, and impaired glucose tolerance: The Australian Diabetes, Obesity, and Lifestyle Study (AusDiab)," *Circulation*, 2007.

[99] G. Roglic *et al.*, "The burden of mortality attributable to diabetes: realistic estimates for the year 2000.," *Diabetes Care*, 2005.

[100] A. D. Association, "Diagnosis and Classification of Diabetes Mellitus DEFINITION AND DESCRIPTION OF DIABETES MELLITUS," *Diabetes Care*, 2014.

[101] E. L. M. Barr, E. J. Boyko, P. Z. Zimmet, R. Wolfe, A. M. Tonkin, and J. E. Shaw, "Continuous relationships between non-diabetic hyperglycaemia and both cardiovascular disease and all-cause mortality: The Australian Diabetes, Obesity, and Lifestyle (AusDiab) study," *Diabetologia*, 2009.

[102] E. B. Levitan, Y. Song, E. S. Ford, and S. Liu, "Is nondiabetic hyperglycemia a risk factor for cardiovascular disease? A meta-analysis of prospective studies," *Arch. Intern. Med.*, 2004.

[103] M. Tominaga, H. Eguchi, H. Manaka, K. Igarashi, T. Kato, and A. Sekikawa, "Impaired glucose tolerance is a risk factor for cardiovascular disease, but not impaired fasting glucose: The Funagata Diabetes Study," *Diabetes Care*, 1999.

[104] J. D. Sorkin, D. C. Muller, J. L. Fleg, and R. Andres, "The relation of fasting and 2-h postchallenge plasma glucose concentrations to mortality: Data from the Baltimore Longitudinal Study of Aging with a critical review of the literature," *Diabetes Care*, 2005.

[105] P. E. Wändell, A. C. Carlsson, and H. Theobald, "The association between BMI value and long-term mortality," *Int. J. Obes.*, 2009.

[106] M. É. Piché, J. F. Arcand-Bossé, J. P. Després, L. Pérusse, S. Lemieux, and S. J. Weisnagel, "What is a normal glucose value? Differences in indexes of plasma glucose homeostasis in subjects with normal fasting glucose," *Diabetes Care*, 2004.

[107] M. A. Westwood *et al.*, "Normalized left ventricular volumes and function in thalassemia major patients with normal myocardial iron," *J. Magn. Reson. Imaging*, 2007.

[108] S. Claster *et al.*, "Nutritional deficiencies in iron overloaded patients with hemoglobinopathies," *Am. J. Hematol.*, 2009.

[109] J. C. Wood *et al.*, "Vitamin D deficiency, cardiac iron and cardiac function in thalassaemia major," *Br. J. Haematol.*, 2008.

[110] A. El-Beshlawy *et al.*, "Improvement of cardiac function in thalassemia major treated with L-carnitine," *Acta Haematol.*, 2004.

[111] J. C. Wood *et al.*, "Physiology and pathophysiology of iron cardiomyopathy in thalassemia," in *Annals of the New York Academy of Sciences*, 2005.

[112] U. N. Dulhare and M. Ayesha, "Extraction of action rules for chronic kidney disease using Naïve bayes classifier," in *2016 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2016*, 2017.

[113] T. J. Peter and K. Somasundaram, "Study and Development of Novel Feature Selection Framework for Heart Disease Prediction," *Int. J. Sci. Res. Publ.*, 2012.

[114] B. Xue, M. Zhang, and W. N. Browne, "Particle swarm optimization for feature selection in classification: A multi-objective approach," *IEEE Trans. Cybern.*, 2013.

[115] Ks. BKavihta Rani AGovrdhan Associate Professor, P. of CSE, and N. Karimnagar Jagtial, "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks," 2010.

[116] N. A. Setiawan, P. A. Venkatachalam, and A. F. M. H, "Rule Selection for Coronary Artery Disease Diagnosis Based on Rough Set," *Int. J. Recent Trends Eng. Technol.*, 2009.

[117] F. Thabtah, "A review of associative classification mining," *Knowledge Engineering Review*. 2007.

[118] F. Tao, F. Murtagh, and M. Farid, "Weighted Association Rule Mining using weighted

support and significance framework," 2004.

[119] M. Stensmo and T. J. Sejnowski, "Automated Medical Diagnosis based on Decision Theory and Learning from Cases," *World Congr. Neural Networks*, 1996.

[120] Wenmin Li, Jiawei Han, and Jian Pei, "CMAR: accurate and efficient classification based on multiple class-association rules," 2002.

[121] M. Al Mamun *et al.*, "Emerging Burden of Cardiovascular Diseases in Bangladesh."

[122] M. Joel M. Gore, "Typical Angina vs. Atypical Chest Pain," *NEJM J. Watch*, vol. 2010, Jul. 2010.

[123] J. Constant, "The diagnosis of nonanginal chest pain.," *Keio J. Med.*, vol. 39, no. 3, pp. 187–92, Sep. 1990.

[124] B. Unal, J. A. Critchley, and S. Capewell, "Modelling the decline in coronary heart disease deaths in England and Wales, 1981-2000: comparing contributions from primary prevention and secondary prevention," *BMJ*, vol. 331, no. 7517, p. 614, Sep. 2005.

[125] "Diabetes - Diagnosis and treatment - Mayo Clinic." [Online]. Available: https://www.mayoclinic.org/diseases-conditions/diabetes/diagnosis-treatment/drc-20371451. [Accessed: 28-Apr-2019].

[126] "Angina (Chest Pain) | American Heart Association." [Online]. Available: https://www.heart.org/en/health-topics/heart-attack/angina-chest-pain. [Accessed: 28-Apr-2019].

[127] H. L. Fred, "Atypical chest pain: a typical humpty dumpty coinage.," *Texas Hear. Inst. J.*, vol. 36, no. 5, pp. 373–4, 2009.

[128] J. Robson, L. Ayerbe, R. Mathur, J. Addo, and A. Wragg, "Clinical value of chest pain presentation and prodromes on the assessment of cardiovascular disease: a cohort study.," *BMJ Open*, vol. 5, no. 4, p. e007251, Apr. 2015.

[129] D. Pei *et al.*, "Relationship of Blood Pressure and Cardiovascular Disease Risk Factors in Normotensive Middle-Aged Men," *Medicine (Baltimore).*, vol. 90, no. 5, pp. 344–349, Sep. 2011.

[130] "How to read an Electrocardiogram (ECG). Part One: Basic principles of the ECG. The normal ECG." [Online]. Available:

http://www.southsudanmedicaljournal.com/archive/may-2010/how-to-read-an-electrocardiogram-ecg.-part-one-basic-principles-of-the-ecg.-the-normal-ecg.html. [Accessed: 28-Apr-2019].

[131] R. E. O'Connor *et al.*, "Part 10: acute coronary syndromes: 2010 American Heart Association Guidelines for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care.," *Circulation*, vol. 122, no. 18 Suppl 3, pp. S787-817, Nov. 2010.

[132] M. Kukar, I. Kononenko, C. Grošelj, K. Kralj, and J. Fettich, "Analysing and improving the diagnosis of ischaemic heart disease with machine learning," *Artif. Intell. Med.*, vol. 16, no. 1, pp. 25–50, May 1999.