A Project Report On

HEART DISEASE PREDICTION USING MACHINE LEARNING

Submitted to

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY

Kukatpally, Hyderabad-500085, Telangana, India

In partial fulfilment of the requirement for the award of degree of

BACHELOR OF TECHNOLOGY

In

Information Technology

By

HIMA BINDU [17E31A1211]

YASMEEN BEGUM [17E31A1233]

Under the guidance of

SIDDI SRINIVAS SIR

Assistant Professor

IT Dept

ESTD:2001

Department of Information Technology

MAHAVEER INSTITUTE OF SCIENCE AND TECHNOLOGY

(Affiliated to JNTU Hyderabad, Approved by AICTE)

Vyasapuri, Bandlaguda, Post: Keshavgiri, Hyderabad-500 005

2020-2021

MAHAVEER INSTITUTE OF SCIENCE AND TECHNOLOGY

(Affiliated to JNTU Hyderabad, Approved by AICTE)

Vyasapuri, Bandlaguda, Post: Keshavgiri, Hyderabad-500005

ESTD:2001

CERTIFICATE

This is to certify that this project work report entitled "**HEART DISEASE PREDICTION USING MACHINE LEARNING"** which is being submitted by **HIMA BINDU[17E31A1211],YASMEEN BEGUM[17E31A1233],** in partial fulfilment for the award of the Degree of Bachelor Of Technology in Information Technology , affiliated of Jawaharlal Nehru Technological University, Hyderabad and is a record of the bonafied work carried out by them under our guidance during 2020-2021.

Signature of Project Guide
Mr. Siddi Srinivas
(Assistant Professor)

Signature of Head of the Department
Dr. A. Nanda Gopal Reddy
(Head Of The Department)

External Examiner

Dr. B. Nageshwara Rao
(Principal)

AKNOWLEGEMENT

We would like to express our deep felt appreciation and gratitude to **Mr.SIDDI SRINIVAS** ,Assistant Professor, Department of IT our project guide, for his skilful guidance, constant supervision, timely suggestion, keen interest and encouragement in completing the individual seminar within the stipulated time.

We wish to express our gratitude to **Mr. SIDDI SRINIVAS,** Project Coordinator,who has shown keen interest and even rendered his valuable guidance in terms of suggestions and encouragement extended to us with an immense care and zeal.

We express our profound sense of gratitude to **Dr.A. NANDA GOPAL REDDY**, Head of Department, IT ,who has served as a host of valuable corrections and for providing us time and amenities to complete this project.

We express our thanks to **Dr. B. NAGESHWARA RAO**, Principal of our college and the management of **Mahaveer Institute of Science and Technology** for providing excellent academic environment in the college.

We wish express our gratitude to the **Members of Staff** and all others who helped us in more than one way. We would also like to thank the **Lab assistants and Programmers** for helping us through our project.

HIMA BINDU [17E31A1211]
YASMEEN BEGUM [17E31A1233]

DECLARATION

We hereby declare that the project entitled “**HEART DISEASE PREDICTION USING MACHINE LEARNING”** submitted to partial fulfilment of the requirements for award of the degree of **Bachelor of Technology** at **Mahaveer Institute of Science and Technology,** affiliated to **Jawaharlal Nehru Technology University, Hyderabad** in authentic work and has not been submitted to any university institute for award of any degree.

HIMA BINDU [17E31A1211]

YASMEEN BEGUM [17E31A1233]

CONTENTS

TITLE **PAGE NO**

CHAPTER 5

CHAPTER 6

CHAPTER 7

CHAPTER 8

CHAPTER 9

ABSTRACT

The major killer cause of human death is Heart Disease (HD). Many people die due to this disease. Lots of researchers have been discovering new technologies to prognosticate the disease early before it's too late for helping healthcare as well as people. These processes are still under research phase. Machine Learning (ML) is faster-emerging technology of Artificial Intelligence (AI) that contributes various algorithms for HD. Based on the proposed problem, ML provides different classification algorithms to divine the probability of patient having HD. For predicting HD, a lot of research scholars contributes their effort in this work using various techniques and algorithms such as Decision Tree (DT), Naïve Bayes (NB), Support Vector Machine (SVM), KNN (KNearest Neighbor),Random Forest etc. In order to give some effort on this work, We are using the UCI repository HD dataset to train a model . The dataset contains 303 instances with 14 attributes that help to train a prediction model that will be for prediction. The main aim of this project is to build an efficient prediction model and deploy for prediction of disease. An HDP Model is built by using NB algorithm that provides 88.163% accuracy among others. Each phase is efficiently done. The project is successfully created with help of requirement analysis and project plan, system design, database design, testing plan, identifying features and functionalities, and system validation and deployment. The limitation of this project is to have only predicted the presence of heart disease but not identify which type of HD does have at patient. In future work, we can enhance the project by appending more detail prediction of HD at patient and incorporate with smart wear devices that integrate to Hospital Emergency System.

Keywords: Machine Learning (ML), Decision Tree (DT), Naïve Bayes (NB), K-Nearest Neighbour (KNN), Random Forest (RF),Support Vector Machine (SVM).

LIST OF FIGURES

LIST OF SCREENS

CHAPTER-1

1.1 INTRODUCTION

Heart disease is one of the prevalent disease that can lead to reduce the lifespan of human beings nowadays. Each year 17.5 million people are dying due to heart disease [1]. Life is dependent on component functioning of heart, because heart is necessary part of our body. Heart disease is a disease that affects on the function of heart [2]. An estimate of a person’s risk for coronary heart disease is important for many aspects of health promotion and clinical medicine. A risk prediction model may be obtained through multivariate regression analysis of a longitudinal study [3]. Due to digital technologies are rapidly growing, healthcare centres store huge amount of data in their database that is very complex and challenging to analysis. Data mining techniques and machine learning algorithms play vital roles in analysis of different data in medical centres. The techniques and algorithms can be directly used on a dataset for creating some models or to draw vital conclusions, and inferences from the dataset. Common attributes used for heart disease are Age, Sex, Fasting Blood Pressure, Chest Pain type, Resting ECG(test that measures the electrical activity of the heart), Number of major vessels colored by fluoroscopy, Threst Blood Pressure (high blood pressure), Serum Cholestrol (determine the risk for developing heart disease), Thalach (maximum heart rate achieved), ST depression (finding on an electrocardiogram, trace in the ST segment is abnormally low below the baseline), painloc (chest pain location (substernal=1, otherwise=0)), Fasting blood sugar, Exang (exercise included angina), smoke, Hypertension, Food habits, weight, height and obesity[4]. Table 1 summarizes the most common types of the heart disease as follows.

1.1.1 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 than 8%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.

Table 1 - Different types of heart disease

Arrhythmia	The heart beat is improper whether it may irregular, too slow or too fast.
Cardiac arrest	An unexpected loss of heart function, consciousness and breathing occur suddenly.
Congestive heart failure	The heart does not pump blood as well as it should, it is the condition of chronic.
Congenital heart disease	The heart’s abnormality which develops before birth.
Coronary artery disease	The heart’s major blood vessels can damage or any disease occurs in the blood vessels.
High Blood Pressure	It has a condition that the force of the blood against the artery walls is too high.
Peripheral artery disease	The narrowed blood vessels which reduce flow of blood in the limbs, is the circulatory condition.

Stroke	Interruption of blood supply occur damage to the brain.

Figure 1 depicts the parts of human heart such as Left atrium, Right atrium, Right ventricle, Left ventricle, Aorta, pulmonary vein, Pulmonary valve, Pulmonary artery, Tricuspid valve, Aortic valve, Mitral valve, Superior vena cava and Interior vena cava.

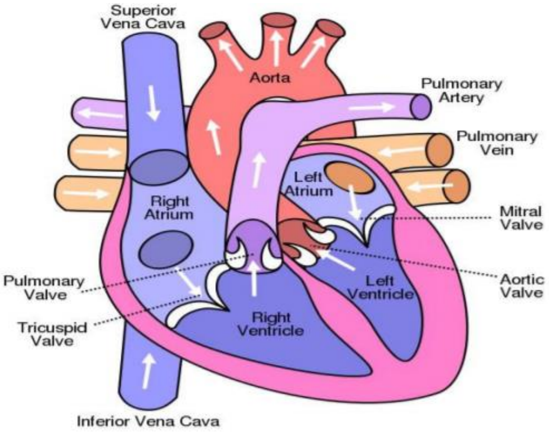
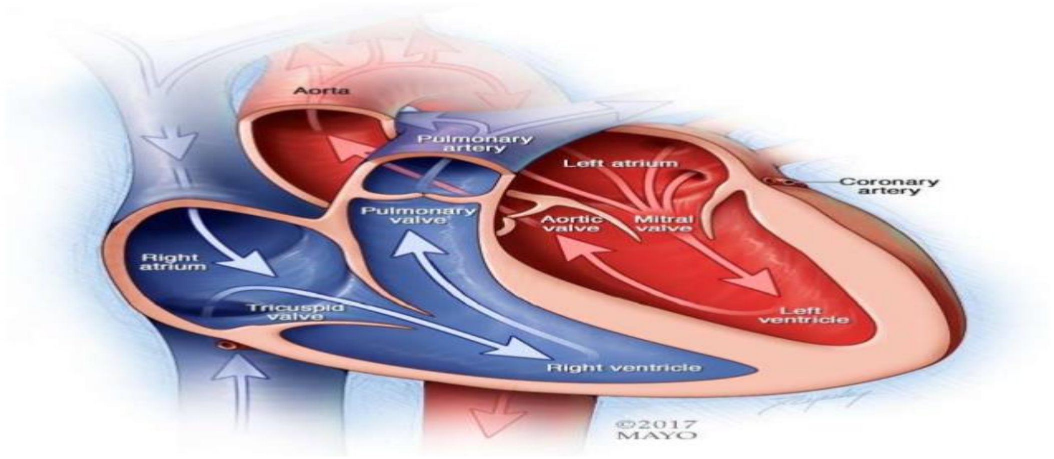


Figure1- Human Heart

The majority of people today experience an unhealthy and fast living style that according to the studies if giving jolt to the heart. The heart is the organ that pumps blood into various parts of the body through the vessels with a proper amount of oxygen and other essential nutrients.The algorithms of machine learning helpful in predicting heart disease

1.1.2 How the Heart works?



1.1.2.1 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.

1.1.2.2 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

1.1.3 Causes of cardiovascular disease

a.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.

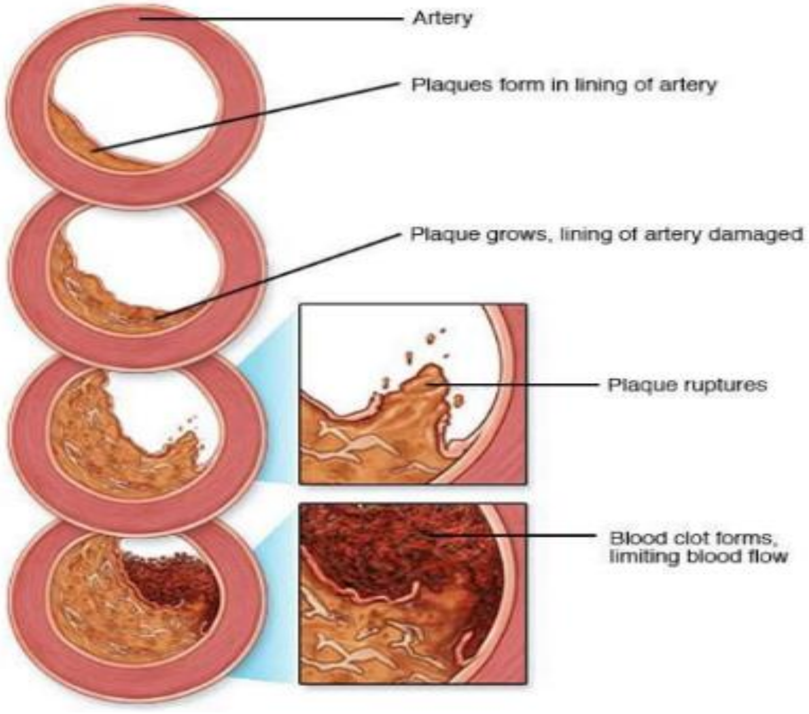

Artery
Plaques form in lining of artery
Plaque grows, lining of artery damaged
Plaque ruptures
Blood clot forms, limiting blood flow

Figure 2-Arteries

b.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.

c.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

d.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. Heart Disease Prediction using Machine Learning Algorithms.

- 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.

1.1.4 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.

1.1.5 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

1.1.6 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.

1.1.6.1 Age

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.

2.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]

3.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 casecontrol 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]

4.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]

5.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]

6.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]

6.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]

1.1.6.2 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]

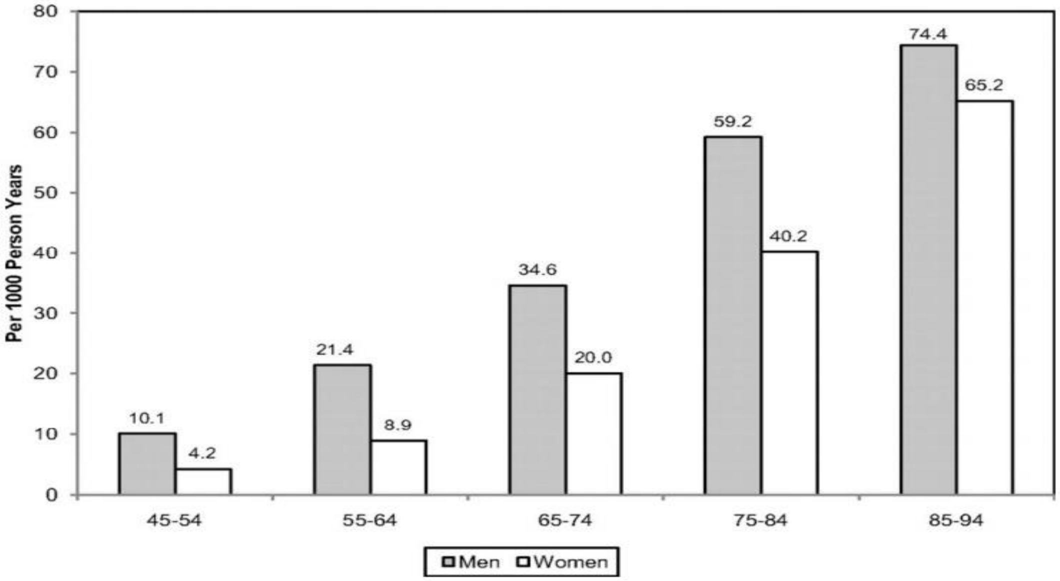


Figure 3- Annual number of adults having diagnosed heart attack or fatal coronary heart disease by age and sex.

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]

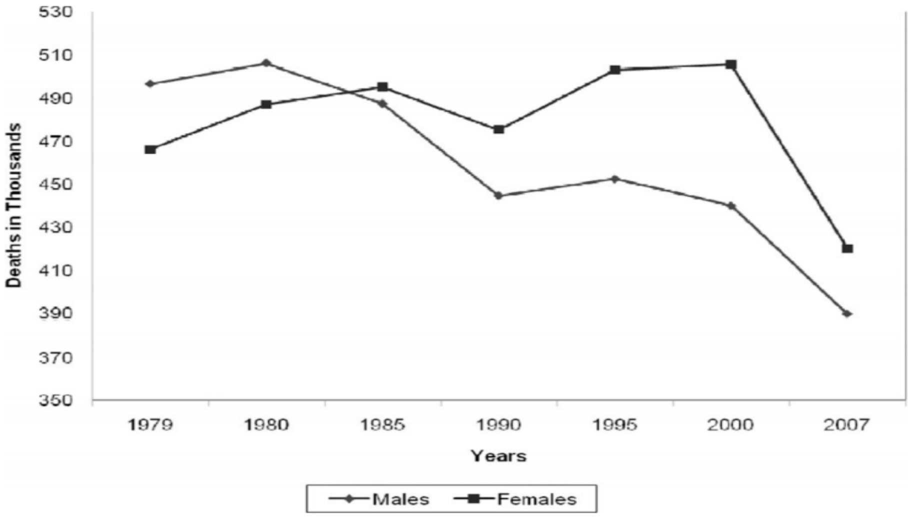


Figure 4 -Trends in the total annual number of deaths caused by cardiovascular disease according to gender

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?

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 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 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. Shaney felt 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]

1.1.6.3 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]

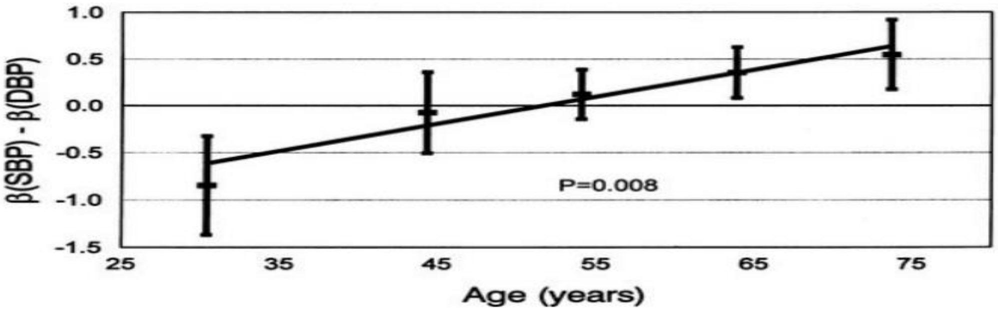


Figure 5 -SBP and DBP wrt to Cardiovascular Disease Presence

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]

1.1.6.4 Chest Pain and its risk factor to Cardiac arrest

Angina is chest pain or discomfort caused when your heart muscle doesn't get enough oxygenrich 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]

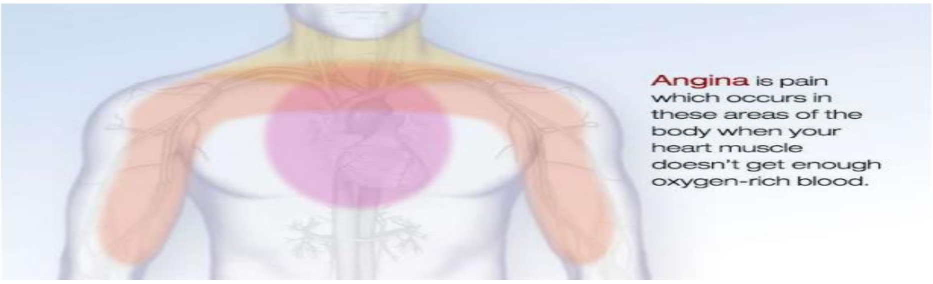

Angina is pain which occurs in these areas of the body when your heart muscle doesn't get enough oxygen-rich blood.

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

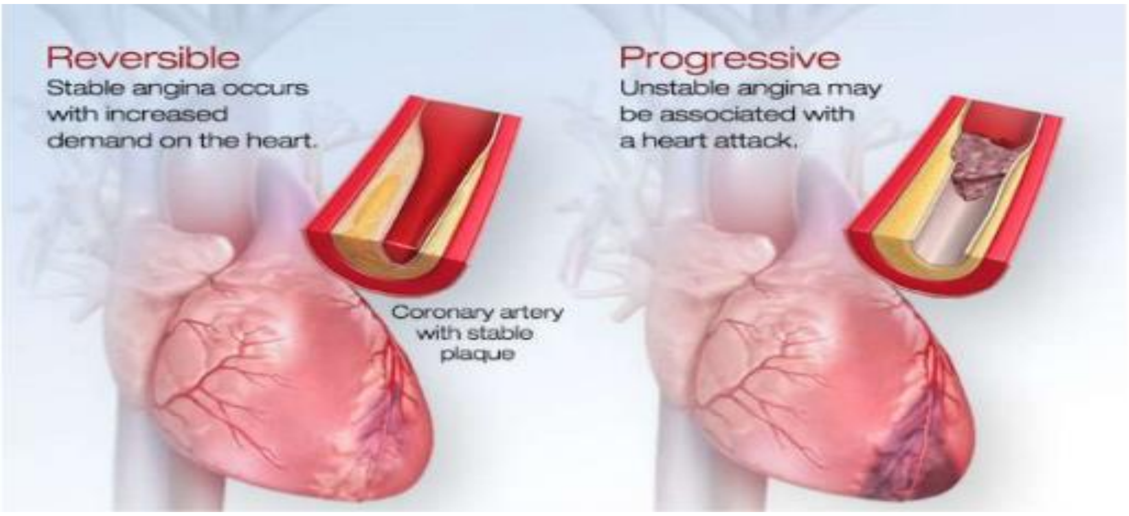

Reversible
Stable angina occurs with increased demand on the heart.
Coronary artery with stable plaque
Progressive
Unstable angina may be associated with a heart attack.

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.

1.1.6.5 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]

1.1.6.5.1 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]

1.1.6.5.2 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]

1.1.6.5.3 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.

1.1.6.6 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 levels had higher levels of arterial stiffness.

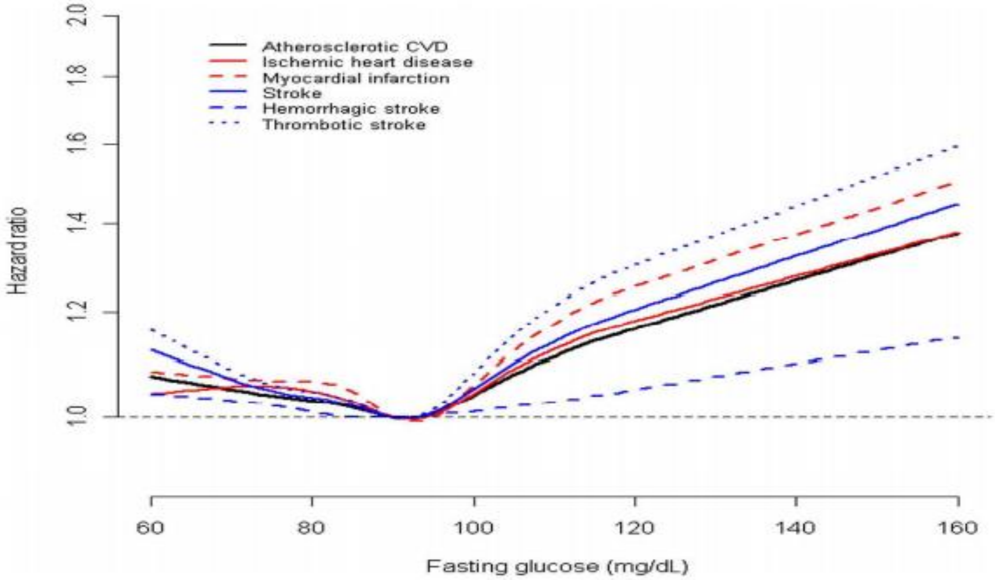


Figure 8 - FBS analysis for Men

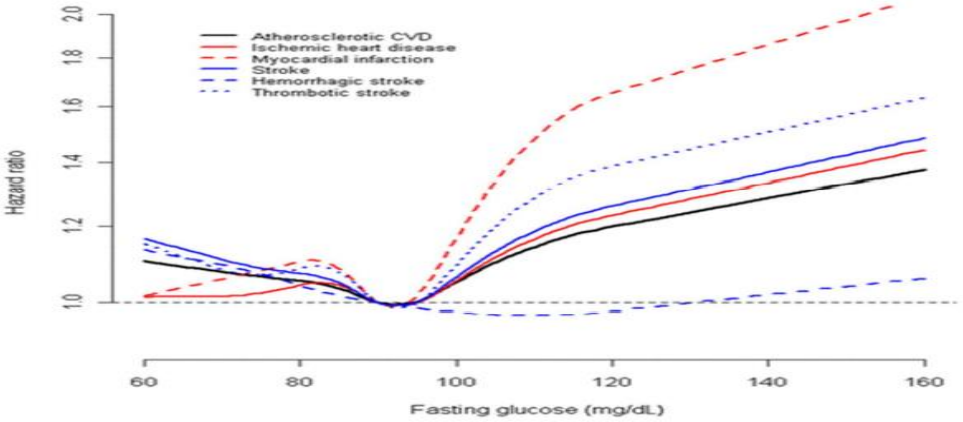


Figure 9 - FBS analysis for Women

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]

1.1.6.7 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]

1.1.6.8 Cardiac Complications in Thalassemia

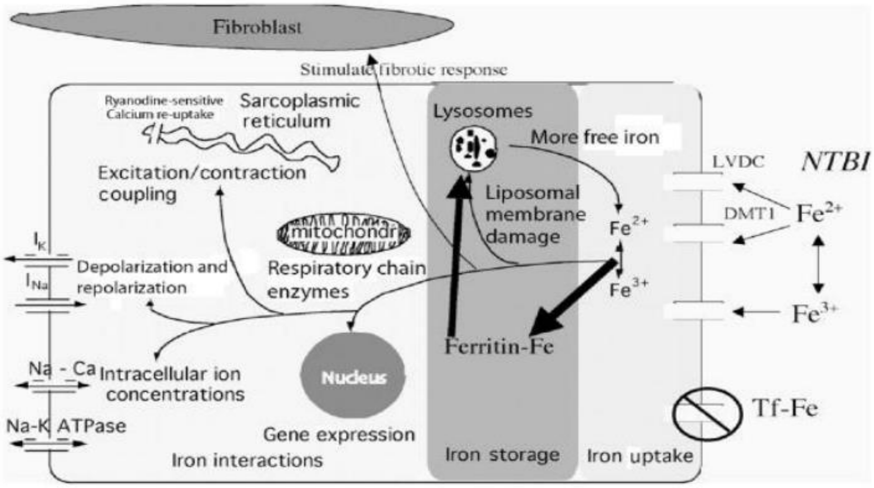
Major Thalassemia major is characterized by chronic ineffective erythropoiesis and anemia as its primary problems. These, in turn, produce physiologic adaptations in the cardiovascular system as well as pathologic/iatrogenic processes such as iron overload, splenectomy, nutritional deficiencies, chronic 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 procoagulants. 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]



<u>Figure 10 - Iron Cardiomyopathy</u> - 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 nontransferrin-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.

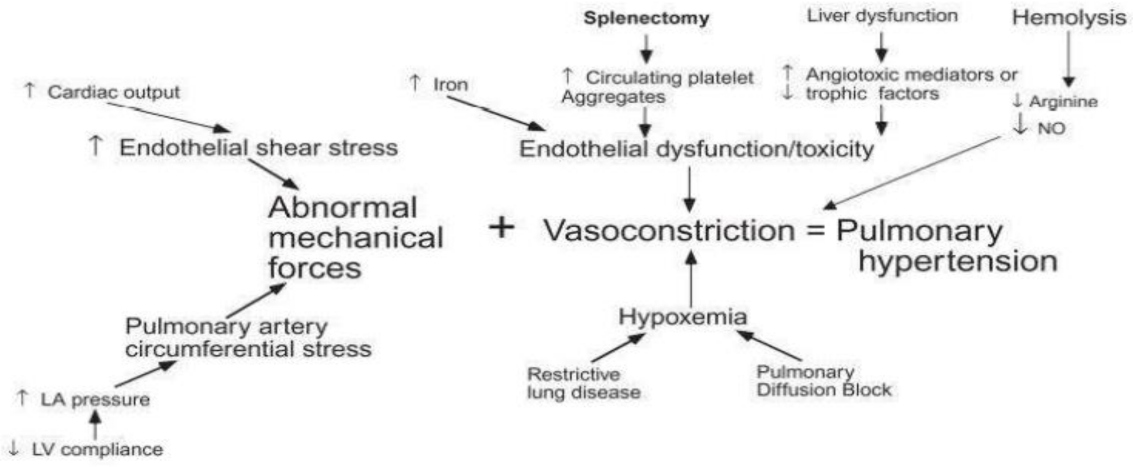

Splenectomy
Liver dysfunction
Hemolysis
↑ Cardiac output
↑ Iron
↑ Circulating platelet Aggregates
↑ Angiotoxic mediators or
↓ trophic factors
↓ Arginine
↓ NO
↑ Endothelial shear stress
Endothelial dysfunction/toxicity
Abnormal mechanical forces
+ Vasoconstriction = Pulmonary hypertension
Pulmonary artery circumferential stress
Hypoxemia
↑ LA pressure
↓ LV compliance
Restrictive lung disease
Pulmonary Diffusion Block

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.

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

1.3 SCOPE

Here the scope of the project is that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions

CHAPTER-2

2.1 LITERATURE REVIEW

According to Senthil Kumar et.al [1], Machine learning (ML) has been shown to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry. ML techniques are being used in recent developments in different areas of the Internet of Things (IoT). Various studies give only a glimpse into predicting heart disease with ML techniques. Authors proposed a novel method that aims at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease. The prediction model is introduced with different combinations of features and several known classification techniques. We produce an enhanced performance level with an accuracy level of 88.7% through the prediction model for heart disease with the hybrid random forest with a linear model (HRFLM).

Anjan Nikhil Repaka, Sai Deepak Ravikanti and Ramya G Franklin[2] used Naive Bayesian method to design and implement heart disease prediction. To achieve this SHDP (Smart Heart Disease Prediction) is built via Navies Bayesian in order to predict risk factors concerning heart disease. The speedy advancement of technology has led to remarkable rise in mobile health technology that being one of the web application. The required data is assembled in a standardized form. For predicting the chances of heart disease in a patient, the following attributes are being fetched from the medical profiles, these include: age, BP, cholesterol, sex, blood sugar etc... The collected attributes acts as input for the Navies Bayesian classification for predicting heart disease. The dataset utilized is split into two sections, 80% dataset is utilized for training and rest 20% is utilized for testing. The proposed approach includes following stages: dataset collection, user registration and login, classification via Navies Bayesian, prediction and secure data transfer by employing Advanced Encryption Standard. Thereafter result is produced. The research elaborates and presents multiple knowledge abstraction techniques by making use of data mining methods which are adopted for heart disease prediction. The output reveals that the established diagnostic system effectively assists in predicting risk factors concerning heart diseases.

According to Ed-Daoudy[3], early detection of heart diseases and continuous monitoring can reduce the mortality rate. The exponential growth of data from different sources such as wearable sensor devices used in Internet of Things health monitoring, streaming system and others have been generating an enormous amount of data on a continuous basis. The combination of streaming big data analytics and machine learning is a breakthrough technology that can have a significant impact in healthcare field especially early detection of heart disease. This technology can be more powerful and less expensive. A real-time heart disease prediction system has been proposed based on apache Spark. The system consists of two main sub parts, namely streaming processing and data storage and visualization. The first uses Spark MLlib with Spark streaming and applies classification model on data events to predict heart disease. The seconds uses Apache Cassandra for storing the large volume of generated data.

Amin Ul Haq et.al[4]have come up with Hybrid Intelligent System Framework for the Prediction of Heart Diseases. The authors asserted that noninvasive-based methods such as machine learning are reliable and efficient. A Machine-learning-based diagnosis system for heart disease prediction by using heart disease dataset was developed using seven popular machine learning algorithms, three feature selection algorithms, the cross-validation method, and seven classifiers performance evaluation metrics such as classification accuracy, specificity, sensitivity, Matthews' correlation coefficient, and execution time. The proposed system can easily identify and classify people with heart disease from healthy people. Additionally, receiver optimistic curves and area under the curves for each classifier was computed. The authors validated performance of the proposed system on full features and on a reduced set of features. The features reduction has an impact on classifiers performance in terms of accuracy and execution time of classifiers.

Md. Shahriare Satu et.al [5] present that Heart Disease is one of the leading diseases that causes enormous loss of lives all over the world. Some unusual approaches to find out significant factors of heart diseases have been considered by the authors. They have used two heart disease data (Cleveland & Hungarian) and both of them are divided into 33%, 65% and 100% data. Values of different range of individual attributes in these data are determined to find out relevant factors of this disease. Then, different semi supervised learning algorithms such as Collective Wrapper, Filtered Collective and Yet Another Semi Supervised Idea are used to analyze heart disease data. Metrics of these classifiers like accuracy, f-measure and area under ROC have been calculated to justify individual classifiers and specify the best semi supervised learning algorithm. This algorithm is explored significant and irrelevant factors of heart disease by removing attributes one after another sequentially and observing the outcomes of classification. Experimental results on two real data demonstrates the effectiveness and efficiency of the analysis.

R. Sharmila et al, [6] suggested using a non-linear heart disease prediction classification algorithm. Big data tools like the Hadoop Distributed File System (HDFS), Mapreduce and SVM are proposed for cardiac disease prediction with optimized attribute description. This thesis explored the use of various methods of data mining for cardiac disease prediction. This proposes to use HDFS to store large amounts of data in different nodes and run the prediction algorithm concurrently using SVM in more than one node. SVM is used in parallel, providing better time than sequential SVM computing.

The prediction and study of Heart Disease Use Techniques for Data Mining was suggested by Chala Beyene et al, [7]. The principal aim is to predict that heart disease will arise in a limited period of time for an automatic early diagnosis of the disease. In the health system with professionals who have no experience and expertise, the suggested approach is also important. It uses various medical features including blood sugar and heart rate, age, sex, some of the features to decide whether you have heart disease. WEKA software is used to measure the performance of data sets.

C. Sowmiya and P.Sumitra[8] believe that It is essential to have a frame work that can effectually recognize the prevalence of heart disease in thousands of samples instantaneously. The authors evaluated the potential of nine (9) classification techniques for prediction of heart disease. Namely decision tree, naïve Bayesian neural network, SVM.ANN, KNN. My proposed algorithm of Apriori algorithm and SVM (support vector machine) in heart disease prediction. Using medical profiles such as a age, sex, blood pressure, chest pain type, fasting blood sugar. It can predict likelihood of patients getting heart disease. Based on this, medical society takes part interest in detecting and preventing the heart disease. From the analysis it have proved that classification based techniques contribute high effectiveness and obtain high accuracy compare than the previous methods.

2.2 .SOFTWARE ENVIRONMENT

2.2.1 MACHINE LEARNING

Machine Learning is a system that can learn from example through self-improvement and without being explicitly coded by programmer. The breakthrough comes with the idea that a machine can singularly learn from the data (i.e., example) to produce accurate results.

Machine learning combines data with statistical tools to predict an output. This output is then used by corporate to makes actionable insights. Machine learning is closely related to data mining and Bayesian predictive modeling. The machine receives data as input, use an algorithm to formulate answers.

A typical machine learning tasks are to provide a recommendation. For those who have a Netflix account, all recommendations of movies or series are based on the user's historical data. Tech companies are using unsupervised learning to improve the user experience with personalizing recommendation.

Machine learning is also used for a variety of task like fraud detection, predictive maintenance, portfolio optimization, automatize task and so on.

2.2.1.1 Machine Learning vs. Traditional Programming-Traditional programming differs significantly from machine learning. In traditional programming, a programmer code all the rules in consultation with an expert in the industry for which software is being developed. Each rule is based on a logical foundation; the machine will execute an output following the logical statement. When the system grows complex, more rules need to be written. It can quickly become unsustainable to maintain.

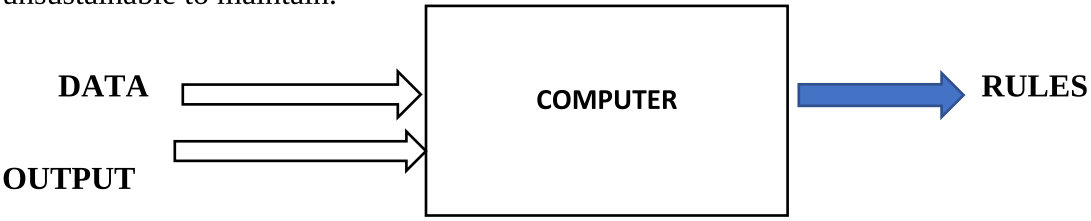


Figure 12-Machine Learning

2.2.1.2 How does Machine learning work?

Machine learning is the brain where all the learning takes place. The way the machine learns is similar to the human being. Humans learn from experience. The more we know, the more easily we can predict. By analogy, when we face an unknown situation, the likelihood of success is lower than the known situation. Machines are trained the same. To make an accurate prediction, the machine sees an example. When we give the machine a similar example, it can figure out the outcome. However, like a human, if it feed a previously unseen example, the machine has difficulties to predict.

The core objective of machine learning is the **learning** and **inference**. First of all, the machine learns through the discovery of patterns. This discovery is made thanks to the **data**. One crucial part of the data scientist is to choose carefully which data to provide to the machine. The list of attributes used to solve a problem is called a **feature vector.** You can think of a feature vector as a subset of data that is used to tackle a problem.

The machine uses some fancy algorithms to simplify the reality and transform this discovery into a **model**. Therefore, the learning stage is used to describe the data and summarize it into a model.

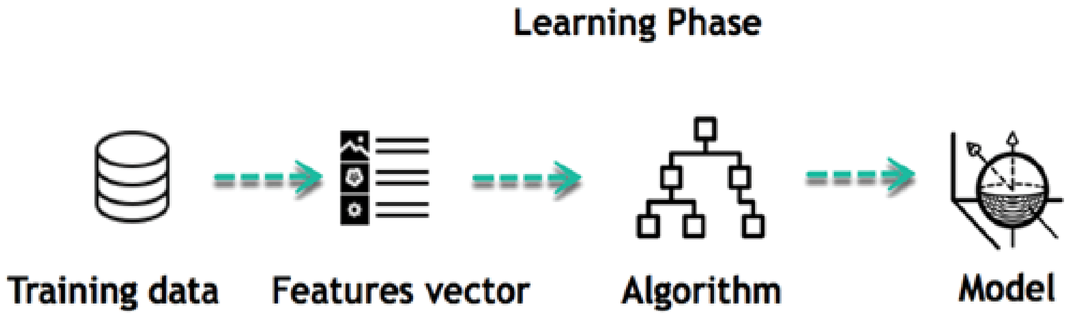


Figure 13– Learning Phase

For instance, the machine is trying to understand the relationship between the wage of an individual and the likelihood to go to a fancy restaurant. It turns out the machine finds a positive relationship between wage and going to a high-end restaurant: This is the model.

2.2.1.3. Inferring

when the model is built, it is possible to test how powerful it is on never-seen-before data. The new data are transformed into a features vector, go through the model and give a prediction. This is all the beautiful part of machine learning. There is no need to update the rules or train the model

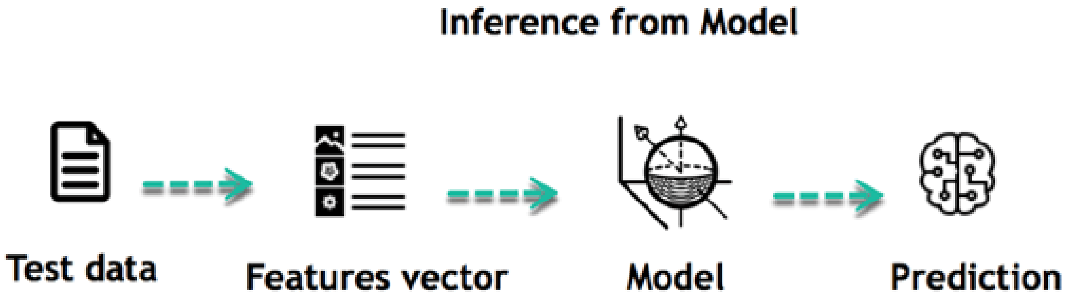


Figure 13- Inference from Model

You can use the model previously trained to make inference on new data. The life of Machine Learning programs is straightforward and can be summarized in the following points:

1. Define a question
2. Collect data
3. Visualize data
4. Train algorithm
5. Test the Algorithm
6. Collect feedback
7. Refine the algorithm
8. Loop 4-7 until the results are satisfying
9. Use the model to make a prediction

Once the algorithm gets good at drawing the right conclusions, it applies that knowledge to new sets of data.

2.2.1.4 Machine learning Algorithms and where they are used?

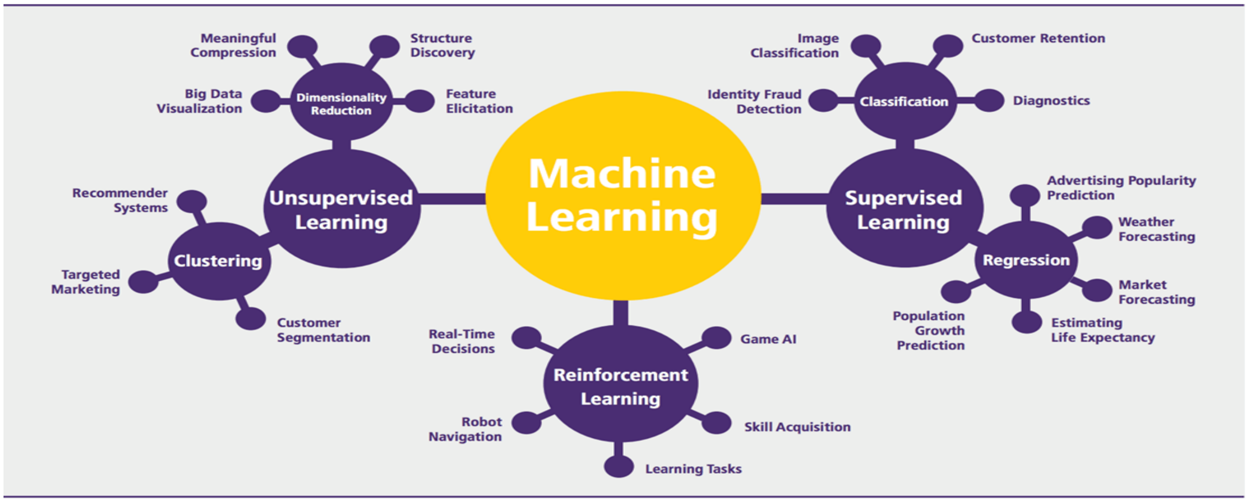


Figure 15- Classification of Machine Learning

Machine learning can be grouped into two broad learning tasks: Supervised and Unsupervised. There are many other algorithms

Supervised learning

An algorithm uses training data and feedback from humans to learn the relationship of given inputs to a given output. For instance, a practitioner can use marketing expense and weather forecast as input data to predict the sales of cans.You can use supervised learning when the output data is known. The algorithm will predict new data.

There are two categories of supervised learning:

- Classification task
- Regression task
- **Classification**

Imagine you want to predict the gender of a customer for a commercial. You will start gathering data on the height, weight, job, salary, purchasing basket, etc. from your customer database. You know the gender of each of your customer, it can only be male or female. The objective of the

Algorithms	Description	Type
Linear regression	Finds a way to correlate each feature to the output to help predict future values.	Regression
Logistic regression	Extension of linear regression that's used for classification tasks. The output variable 3is binary (e.g., only black or white) rather than continuous (e.g., an infinite list of potential colors)	Classification
Decision tree	Highly interpretable classification or regression model that splits data-feature values into branches at decision nodes (e.g., if a feature is a color, each possible color becomes a new branch) until a final decision output is made	Regression Classification
Naive Bayes	The Bayesian method is a classification method that makes use of the Bayesian theorem. The theorem updates the prior knowledge of an event with the independent probability of each feature that can affect the event.	Regression Classification
Support vector machine	Support Vector Machine, or SVM, is typically used for the classification task. SVM algorithm finds a hyperplane that optimally divided the classes. It is best used with a non-linear solver.	Regression (not very common) Classification
Random forest	The algorithm is built upon a decision tree to improve the accuracy drastically. Random forest generates many times simple decision trees and uses the 'majority vote' method to decide on which label to return. For the classification task, the final prediction will be the one with the most vote; while for the regression task, the average prediction of all the trees is the final prediction.	Regression Classification
AdaBoost	Classification or regression technique that uses a multitude of models to come up with a decision but weighs them based on their accuracy in predicting the outcome	Regression Classification
Gradient-boosting trees	Gradient-boosting trees is a state-of-the-art classification/regression technique. It is focusing on the error committed by the previous trees and tries to correct it.	Regression Classification

classifier will be to assign a probability of being a male or a female (i.e., the label) based on the information (i.e., features you have collected). When the model learned how to recognize male or female, you can use new data to make a prediction. For instance, you just got new information from an unknown customer, and you want to know if it is a male or female. If the classifier predicts male = 70%, it means the algorithm is sure at 70% that this customer is a male, and 30% it is a female.

- The label can be of two or more classes. The above example has only two classes, but if a classifier needs to predict object, it has dozens of classes (e.g., glass, table, shoes, etc. each object represents a class)

Regression-

Algorithm	**Description**	**Type**
K-means clustering	Puts data into some groups (k) that each contains data with similar characteristics (as determined by the model, not in advance by humans)	Clustering
Gaussian mixture model	A generalization of k-means clustering that provides more flexibility in the size and shape of groups (clusters	Clustering
Hierarchical clustering	Splits clusters along a hierarchical tree to form a classification system. Can be used for Cluster loyalty-card customer	Clustering
Recommender system	Help to define the relevant data for making a recommendation.	Clustering
PCA/T-SNE	Mostly used to decrease the dimensionality of the data. The algorithms reduce the number of features to 3 or 4 vectors with the highest variances.	Dimension Reduction

When the output is a continuous value, the task is a regression. For instance, a financial analyst may need to forecast the value of a stock based on a range of feature like equity, previous stock performances, macroeconomics index. The system will be trained to estimate the price of the stocks with the lowest possible error.

Unsupervised learning

In unsupervised learning, an algorithm explores input data without being given an explicit output variable (e.g., explores customer demographic data to identify patterns)

You can use it when you do not know how to classify the data, and you want the algorithm to find patterns and classify the data for you

2.2.1.5 Application of Machine learning

Augmentation:

- Machine learning, which assists humans with their day-to-day tasks, personally or commercially without having complete control of the output. Such machine learning is used in different ways such as Virtual Assistant, Data analysis, software solutions. The primary user is to reduce errors due to human bias.

Automation:

- Machine learning, which works entirely autonomously in any field without the need for any human intervention. For example, robots performing the essential process steps in manufacturing plants.

Finance Industry

- Machine learning is growing in popularity in the finance industry. Banks are mainly using ML to find patterns inside the data but also to prevent fraud.

Government organization

- The government makes use of ML to manage public safety and utilities. Take the example of China with the massive face recognition. The government uses Artificial intelligence to prevent jaywalker.

Healthcare industry

- Healthcare was one of the first industry to use machine learning with image detection.

Marketing

- Broad use of AI is done in marketing thanks to abundant access to data. Before the age of mass data, researchers develop advanced mathematical tools like Bayesian analysis to estimate the value of a customer. With the boom of data, marketing department relies on AI to optimize the customer relationship and marketing campaign.

Example of application of Machine Learning in Supply Chain

Machine learning gives terrific results for visual pattern recognition, opening up many potential applications in physical inspection and maintenance across the entire supply chain network.

Unsupervised learning can quickly search for comparable patterns in the diverse dataset. In turn, the machine can perform quality inspection throughout the logistics hub, shipment with damage and wear.

For instance, IBM's Watson platform can determine shipping container damage. Watson combines visual and systems-based data to track, report and make recommendations in real-time.

In past year stock manager relies extensively on the primary method to evaluate and forecast the inventory. When combining big data and machine learning, better forecasting techniques have been implemented (an improvement of 20 to 30 % over traditional forecasting tools). In term of sales, it means an increase of 2 to 3 % due to the potential reduction in inventory costs.

Example of Machine Learning Google Car

For example, everybody knows the Google car. The car is full of lasers on the roof which are telling it where it is regarding the surrounding area. It has radar in the front, which is informing the car of the speed and motion of all the cars around it. It uses all of that data to figure out not only how to drive the car but also to figure out and predict what potential drivers around the car are going to do. What's impressive is that the car is processing almost a gigabyte a second of data.

2.2.1.6 Deep Learning

Deep learning is a computer software that mimics the network of neurons in a brain. It is a subset of machine learning and is called deep learning because it makes use of deep neural networks. The machine uses different layers to learn from the data. The depth of the model is represented by the number of layers in the model. Deep learning is the new state of the art in term of AI. In deep learning, the learning phase is done through a neural network.

2.2.1.7 Reinforcement Learning

Reinforcement learning is a subfield of machine learning in which systems are trained by receiving virtual "rewards" or "punishments," essentially learning by trial and error. Google's DeepMind has used reinforcement learning to beat a human champion in the Go games. Reinforcement learning is also used in video games to improve the gaming experience by providing smarter bot.

One of the most famous algorithms are:

- Q-learning
- Deep Q network
- State-Action-Reward-State-Action (SARSA)
- Deep Deterministic Policy Gradient (DDPG)

2.2.1.8 Applications/ Examples of deep learning applications

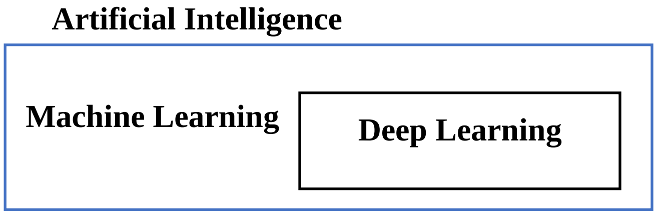
AI in Finance: The financial technology sector has already started using AI to save time, reduce costs, and add value. Deep learning is changing the lending industry by using more robust credit scoring. Credit decision-makers can use AI for robust credit lending applications to achieve faster, more accurate risk assessment, using machine intelligence to factor in the character and capacity of applicants.

Underwrite is a Fintech company providing an AI solution for credit makers company. underwrite.ai uses AI to detect which applicant is more likely to pay back a loan. Their approach radically outperforms traditional methods.

AI in HR: Under Armour, a sportswear company revolutionizes hiring and modernizes the candidate experience with the help of AI. In fact, Under Armour Reduces hiring time for its retail stores by 35%. Under Armour faced a growing popularity interest back in 2012. They had, on average, 30000 resumes a month. Reading all of those applications and begin to start the screening and interview process was taking too long. The lengthy process to get people hired and on-boarded impacted Under Armour's ability to have their retail stores fully staffed, ramped and ready to operate.

At that time, Under Armour had all of the 'must have' HR technology in place such as transactional solutions for sourcing, applying, tracking and onboarding but those tools weren't useful enough. Under armour choose **HireVue**, an AI provider for HR solution, for both on-demand and live interviews. The results were bluffing; they managed to decrease by 35% the time to fill. In return, the hired higher quality staffs.

AI in Marketing: AI is a valuable tool for customer service management and personalization challenges. Improved speech recognition in call-center management and call routing as a result of the aplpication of AI techniques allows a more seamless experience for customers.For example, deep-learning analysis of audio allows systems to assess a customer's emotional tone. If the customer is responding poorly to the AI chatbot, the system can be rerouted the conversation to real, human operators that take over the issue.



2.2.1.9 Difference between Machine Learning and Deep Learning

	Machine learning	Deep learning
Training dataset	Small	Large
Choose features	Yes	No
Number of algorithms	Many	Few
Training time	Short	Long

When to use ML or DL? In the table below, we summarize the difference between machine learning and deep learning.

	Machine Learning	**Deep Learning**
Data Dependencies	Excellent performances on a small/medium dataset	Excellent performance on a big dataset
Hardware dependencies	Work on a low-end machine.	Requires powerful machine, preferably with GPU: DL performs a significant amount of matrix multiplication
Feature engineering	Need to understand the features that represent the data	No need to understand the best feature that represents the data
Execution time	From few minutes to hours	Up to weeks. Neural Network needs to compute a significant number of weights
Interpretability	Some algorithms are easy to interpret (logistic, decision tree), some are almost impossible(SVM, XGBoost)	Difficult to impossible

With machine learning, you need fewer data to train the algorithm than deep learning. Deep learning requires an extensive and diverse set of data to identify the underlying structure. Besides, machine learning provides a faster-trained model. Most advanced deep learning architecture can take days to a week to train. The advantage of deep learning over machine learning is it is highly accurate. You do not need to understand what features are the best representation of the data; the neural network learned how to select critical features. In machine learning, you need to choose for yourself what features to include in the model

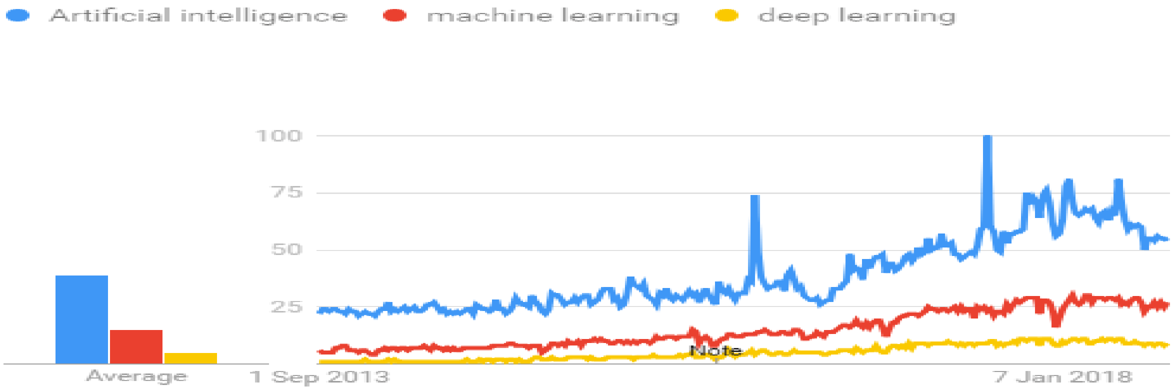


Figure 16 – Statistics of Artificial Intelligence,Machine Learning,Deep Learning

2.2.1.10 TensorFlow -the most famous deep learning library in the world is Google's TensorFlow. Google product uses machine learning in all of its products to improve the search engine, translation, image captioning or recommendations.

To give a concrete example, Google users can experience a faster and more refined the search with AI. If the user types a keyword a the search bar, Google provides a recommendation about what could be the next word.

Google wants to use machine learning to take advantage of their massive datasets to give users the best experience. Three different groups use machine learning:

- Researchers
- Data scientists
- Programmers.

They can all use the same toolset to collaborate with each other and improve their efficiency.

Google does not just have any data; they have the world's most massive computer, so TensorFlow was built to scale. TensorFlow is a library developed by the Google Brain Team to accelerate machine learning and deep neural network research.It was built to run on multiple CPUs or GPUs and even mobile operating systems, and it has several wrappers in several languages like Python, C++ or Java.

2.2.1.11 TensorFlow Architecture

Tensor flow architecture works in three parts:

- Pre processing the data
- Build the model
- Train and estimate the model

It is called Tensor flow because it takes input as a multi-dimensional array, also known as **tensors**. You can construct a sort of **flowchart** of operations (called a Graph) that you want to perform on that input. The input goes in at one end, and then it flows through this system of multiple operations and comes out the other end as output. This is why it is called TensorFlow because the tensor goes in it flows through a list of operations, and then it comes out the other side.

2.2.1.12 Where can Tensor flow run?

TensorFlow can hardware, and software requirements can be classified into

Development Phase: This is when you train the mode. Training is usually done on your Desktop or laptop.

Run Phase or Inference Phase: Once training is done Tensorflow can be run on many different platforms. You can run it on

- Desktop running Windows, macOS or Linux
- Cloud as a web service
- Mobile devices like iOS and Android.
- You can train it on multiple machines then you can run it on a different machine, once you have the trained model.

The model can be trained and used on GPUs as well as CPUs. GPUs were initially designed for video games. In late 2010, Stanford researchers found that GPU was also very good at matrix operations and algebra so that it makes them very fast for doing these kinds of calculations. Deep learning relies on a lot of matrix multiplication. TensorFlow is very fast at computing the matrix multiplication because it is written in C++. Although it is implemented in C++, TensorFlow can be accessed and controlled by other languages mainly, Python.

Finally, a significant feature of Tensor Flow is the Tensor Board. The Tensor Board enables to monitor graphically and visually what TensorFlow is doing.

2.2.1.13 List of Prominent Algorithms supported by TensorFlow

- Linear regression: tf. estimator .Linear Regressor
- Classification :tf. Estimator .Linear Classifier
- Deep learning classification: tf. estimator. DNN Classifier
- Booster tree regression: tf.estimator.BoostedTreesRegressor
- Boosted tree classification: tf.estimator.BoostedTreesClassifier

2.2.2 PYTHON

Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

- **Python is Interpreted:** Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.

- **Python is Interactive:** You can actually sit at a Python prompt and interact with the interpreter directly to write your programs.

- **Python is Object-Oriented:** Python supports Object-Oriented style or technique of programming that encapsulates code within objects.

- **Python is a Beginner's Language:** Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

2.2.2.1 History of Python

Python was developed by Guido van Rossum in the late eighties and early nineties at the National Research Institute for Mathematics and Computer Science in the Netherlands.

Python is derived from many other languages, including ABC, Modula-3, C, C++, Algol-68, SmallTalk, Unix shell, and other scripting languages.

Python is copyrighted. Like Perl, Python source code is now available under the GNU General Public License (GPL).

Python is now maintained by a core development team at the institute, although Guido van Rossum still holds a vital role in directing its progress.

2.2.2.2 Python Features

Python's features include:

- **Easy-to-learn:** Python has few keywords, simple structure, and a clearly defined syntax. This allows the student to pick up the language quickly.
- **Easy-to-read:** Python code is more clearly defined and visible to the eyes.
- **Easy-to-maintain:** Python's source code is fairly easy-to-maintain.
- **A broad standard library:** Python's bulk of the library is very portable and cross-platform compatible on UNIX, Windows, and Macintosh.
- **Interactive Mode:** Python has support for an interactive mode which allows interactive testing and debugging of snippets of code.
- **Portable:** Python can run on a wide variety of hardware platforms and has the same interface on all platforms.
- **Extendable:** You can add low-level modules to the Python interpreter. These modules enable programmers to add to or customize their tools to be more efficient.
- **Databases:** Python provides interfaces to all major commercial databases.

- **GUI Programming:** Python supports GUI applications that can be created and ported to many system calls, libraries, and windows systems, such as Windows MFC, Macintosh, and the X Window system of Unix.
- **Scalable:** Python provides a better structure and support for large programs than shell scripting.

Apart from the above-mentioned features, Python has a big list of good features, few are listed below:

- IT supports functional and structured programming methods as well as OOP.
- It can be used as a scripting language or can be compiled to byte-code for building large applications.
- It provides very high-level dynamic data types and supports dynamic type checking.
- IT supports automatic garbage collection.
- It can be easily integrated with C, C++, COM, ActiveX, CORBA, and Java.

Python is available on a wide variety of platforms including Linux and Mac OS X. Let's understand how to set up our Python environment.

2.2.2.3 PYTHON ENVIRONMENT

Python is available on a wide variety of platforms including Linux and Mac OS X. Let's understand how to set up our Python environment.

2.2.2.4 Python’s standard library

- Pandas
- Numpy
- Sklearn
- seaborn
- matplotlib
- Importing Datasets

PANDAS

Pandas is quite a game changer when it comes to analyzing data with Python and it is one of the most preferred and widely used tools in data munging/wrangling if not the most used one. Pandas is an open source.

What’s cool about Pandas is that it takes data (like a CSV or TSV file, or a SQL database) and creates a Python object with rows and columns called data frame that looks very similar to table in a statistical software (think Excel or SPSS for example. People who are familiar with R would see similarities to R too). This is so much easier to work with in comparison to working with lists and/or dictionaries through for loops or list comprehension.

1. Installation and Getting Started

In order to “get” Pandas you would need to install it. You would also need to have Python 2.7 and above as a pre-requirement for installation. It is also dependent on other libraries (like NumPy) and has optional dependancies (like Matplotlib for plotting). Therefore, I think that the easiest way to get Pandas set up is to install it through a package like the Anaconda distribution , “a cross platform distribution for data analysis and scientific computing.”

In order to use Pandas in your Python IDE (Integrated Development Environment) like Jupyter Notebook or Spyder (both of them come with Anaconda by default), you need to import the Pandas library first. Importing a library means loading it into the memory and then it’s there for you to work with. In order to import Pandas all you have to do is run the following code:

- **import pandas as pd**
- **import numpy as np**

Usually you would add the second part (‘as pd’) so you can access Pandas with ‘pd.command’ instead of needing to write ‘pandas.command’ every time you need to use it. Also, you would import numpy as well, because it is very useful library for scientific computing with Python. Now Pandas is ready for use! Remember, you would need to do it every time you start a new Jupyter Notebook, Spyder file etc.

2.Working with Pandas

Loading and Saving Data with Pandas

When you want to use Pandas for data analysis, you'll usually use it in one of three different ways:

- Convert a Python's list, dictionary or Numpy array to a Pandas data frame
- Open a local file using Pandas, usually a CSV file, but could also be a delimited text file (like TSV), Excel, etc
- Open a remote file or database like a CSV or a JSONon a website through a URL or read from a SQL table/database

There are different commands to each of these options, but when you open a file, they would look like this:

- **pd.read_filetype()**

As I mentioned before, there are different filetypes Pandas can work with, so you would replace "filetype" with the actual, well, filetype (like CSV). You would give the path, filename etc inside the parenthesis. Inside the parenthesis you can also pass different arguments that relate to how to open the file. There are numerous arguments and in order to know all you them, you would have to read the documentation (for example, the documentation for pd.read_csv() would contain all the arguments you can pass in this Pandas command).

In order to convert a certain Python object (dictionary, lists etc) the basic command is:

- **pd.DataFrame()**

Inside the parenthesis you would specify the object(s) you're creating the data frame from. This command also has different arguments .

Youcan also save a data frame you're working with/on to different kinds of files (like CSV, Excel, JSON and SQL tables). The general code for that is:

- **df.to_filetype(filename)**

3.Viewing and Inspecting Data

Now that you’ve loaded your data, it’s time to take a look. How does the data frame look? Running the name of the data frame would give you the entire table, but you can also get the first n rows with df.head(n) or the last n rows with df.tail(n). df.shape would give you the number of rows and columns. df.info() would give you the index, datatype and memory information. The command s.value_counts(dropna=False) would allow you to view unique values and counts for a series (like a column or a few columns). A very useful command is df.describe() which inputs summary statistics for numerical columns. It is also possible to get statistics on the entire data frame or a series (a column etc):

- df.mean() Returns the mean of all columns
- df.corr() Returns the correlation between columns in a data frame
- df.count() Returns the number of non-null values in each data frame column
- df.max()Returns the highest value in each column
- df.min()Returns the lowest value in each column
- df.median()Returns the median of each column
- df.std()Returns the standard deviation of each column

5. Selection of Data

One of the things that is so much easier in Pandas is selecting the data you want in comparison to selecting a value from a list or a dictionary. You can select a column (df[col]) and return column with label col as Series or a few columns (df[[col1, col2]]) and returns columns as a new DataFrame. You can select by position (s.iloc[0]), or by index (s.loc['index_one']) . In order to select the first row you can use df.iloc[0,:] and in order to select the first element of the first column you would run df.iloc[0,0] . These can also be used in different combinations, so I hope it gives you an idea of the different selection and indexing you can perform in Pandas.

6.Filter, Sort and Groupby

You can use different conditions to filter columns. For example, df[df[year] > 1984] would give you only the column year is greater than 1984. You can use & (and) or | (or) to add different conditions to your filtering. This is also called boolean filtering.

It is possible to sort values in a certain column in an ascending order using df.sort_values(col1) ; and also in a descending order using df.sort_values(col2,ascending=False). Furthermore, it's possible to sort values by col1 in ascending order then col2 in descending order by using df.sort_values([col1,col2],ascending=[True,False]).

The last command in this section is groupby. It involves splitting the data into groups based on some criteria, applying a function to each group independently and combining the results into a data structure. df.groupby(col) returns a groupby object for values from one column while df.groupby([col1,col2]) returns a groupby object for values from multiple columns.

7.Data Cleaning

Data cleaning is a very important step in data analysis. For example, we always check for missing values in the data by running pd.isnull() which checks for null Values, and returns a boolean array (an array of true for missing values and false for non-missing values). In order to get a sum of null/missing values, run pd.isnull().sum(). pd.notnull() is the opposite of pd.isnull(). After you get a list of missing values you can get rid of them, or drop them by using df.dropna() to drop the rows or df.dropna(axis=1) to drop the columns. A different approach would be to fill the missing values with other values by using df.fillna(x) which fills the missing values with x (you can put there whatever you want) or s.fillna(s.mean()) to replace all null values with the mean (mean can be replaced with almost any function from the statistics section).

It is sometimes necessary to replace values with different values. For example, s.replace(1,'one') would replace all values equal to 1 with 'one'. It's possible to do it for multiple values: s.replace([1,3],['one','three'])would replace all 1 with 'one' and 3 with 'three'. You can also rename specific columns by running: df.rename(columns={'old_name': 'new_ name'})or use df.set_index('column_one') to change the index of the data frame.

8. Join/Combine

The last set of basic Pandas commands are for joining or combining data frames or rows/columns. The three commands are:

- df1.append(df2)— add the rows in df1 to the end of df2 (columns should be identical)
- df.concat([df1, df2],axis=1)—add the columns in df1 to the end of df2 (rows should be identical)

- df1.join(df2,on=col1,how='inner')—SQL-style join the columns in df1with the columns on df2 where the rows for colhave identical values. how can be equal to one of: 'left', 'right', 'outer', 'inner'

B.NUMPY

Numpy is one such powerful library for array processing along with a large collection of high-level mathematical functions to operate on these arrays. These functions fall into categories like Linear Algebra, Trigonometry, Statistics, Matrix manipulation, etc.

1.**Getting NumPy**

NumPy's main object is a homogeneous multidimensional array. Unlike python's array class which only handles one-dimensional array, NumPy's ndarray class can handle multidimensional array and provides more functionality. NumPy's dimensions are known as axes. For example, the array below has 2 dimensions or 2 axes namely rows and columns. Sometimes dimension is also known as a rank of that particular array or matrix.

2.Importing NumPy

NumPy is imported using the following command. Note here np is the convention followed for the alias so that we don't need to write numpy every time.

- import numpy as np

NumPy is the basic library for scientific computations in Python and this article illustrates some of its most frequently used functions. Understanding NumPy is the first major step in the journey of machine learning and deep learning.

3. Sklearn

In python, scikit-learn library has a pre-built functionality under sklearn. Pre processing.

Next thing is to do feature extraction Feature extraction is an attribute reduction process. Unlike feature selection, which ranks the existing attributes according to their predictive significance, feature extraction actually transforms the attributes. The transformed attributes, or features, are linear combinations of the original attributes. Finally our models are trained using Classifier algorithm.. We use nltk . classify module on Natural Language Toolkit library on Python. We use the labelled dataset gathered . The rest of our labelled data will be used to evaluate the models. Some machine learning algorithms were used to classify pre processed data. The chosen classifiers were Decision tree , Support Vector Machines and Random forest. These algorithms are very popular in text classification tasks.

4.SEABORN

3.2.3 Data Visualization in Python

Data visualization is the discipline of trying to understand data by placing it in a visual context, so that patterns, trends and correlations that might not otherwise be detected can be exposed.

Python offers multiple great graphing libraries that come packed with lots of different features. No matter if you want to create interactive, live or highly customized plots python has a excellent library for you.

3.2.4 To get a little overview here are a few popular plotting libraries:

- **Matplotlib:** low level, provides lots of freedom
- **Pandas Visualization:** easy to use interface, built on Matplotlib
- **Seaborn:** high-level interface, great default styles
- **ggplot:** based on R’s ggplot2, uses Grammar of Graphics
- **Plotly:** can create interactive plots

In this article, we will learn how to create basic plots using Matplotlib, Pandas visualization and Seaborn as well as how to use some specific features of each library. This article will focus on the syntax and not on interpreting the graphs.

1.Matplotlib

Matplotlib is the most popular python plotting library. It is a low level library with a Matlab like interface which offers lots of freedom at the cost of having to write more code.

1. To install Matplotlib pip and conda can be used.
2. pip install matplotlib
3. conda install matplotlib

Matplotlib is specifically good for creating basic graphs like line charts, bar charts, histograms and many more. It can be imported by typing:

- **import matplotlib.pyplot as plt**

2.Line Chart

In Matplotlib we can create a line chart by calling the `plot` method. We can also plot multiple columns in one graph, by looping through the columns we want, and plotting each column on the same axis.

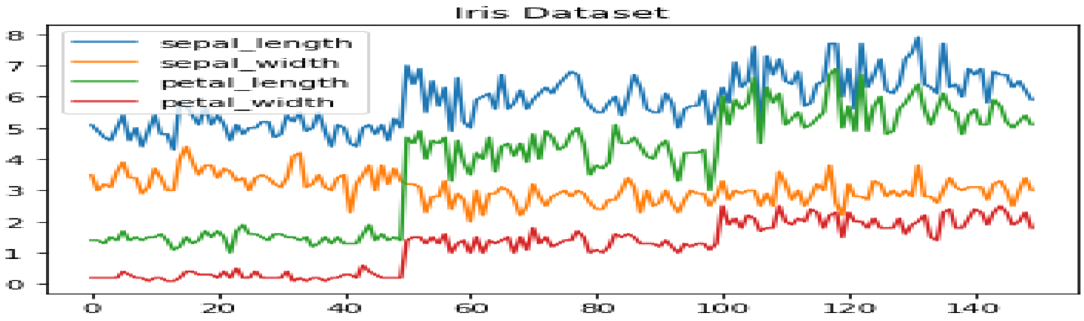


Figure 17-Line Chart

3.Histogram

In Matplotlib we can create a Histogram using the `hist` method. If we pass it categorical data like the points column from the wine-review dataset it will automatically calculate how often each class occurs.

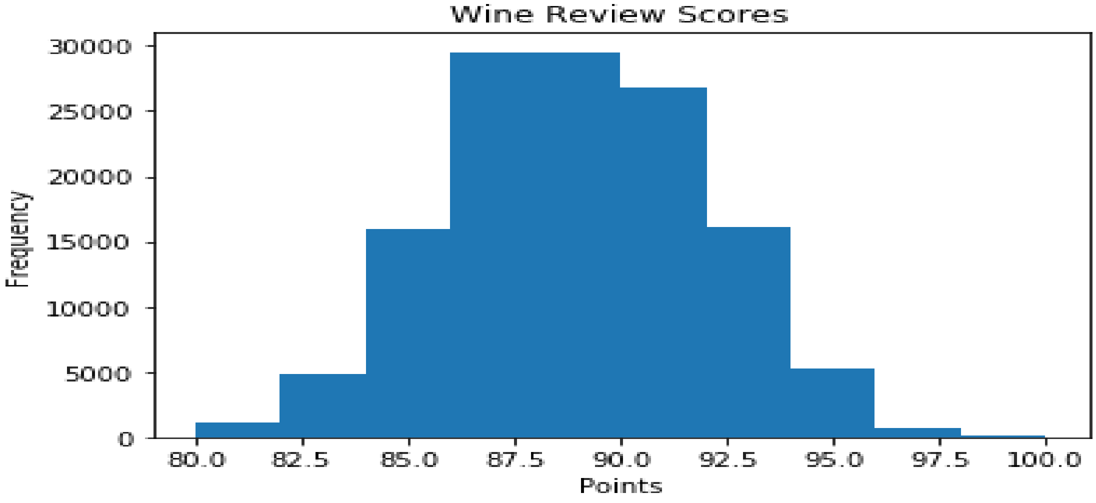


Figure 18-Histogram

4.Bar Chart

A bar-chart can be created using the `bar` method. The bar-chart isn't automatically calculating the frequency of a category so we are going to use pandas `value_counts` function to do this. The bar-chart is useful for categorical data that doesn’t have a lot of different categories (less than 30) because else it can get quite messy.

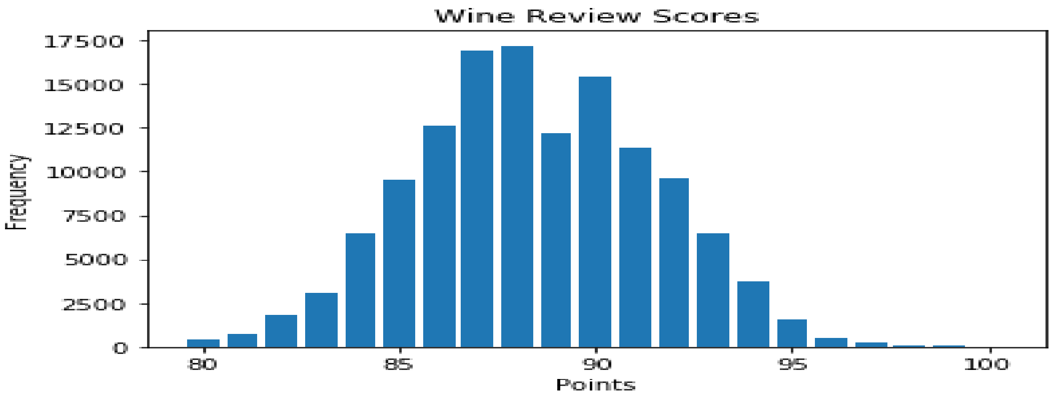


Figure 19-Bar-Chart

Pandas Visualization

Pandas is a open source high-performance, easy-to-use library providing data structures, such as dataframes, and data analysis tools like the visualization tools we will use in this article.

Pandas Visualization makes it really easy to create plots out of a pandas dataframe and series. It also has a higher level API than Matplotlib and therefore we need less code for the same results.

1. **Pandas can be installed using either pip or conda.**
2. **pip install pandas**
3. **conda install pandas**

5.Heatmap

A Heatmap is a graphical representation of data where the individual values contained in a matrix are represented as colors. Heatmaps are perfect for exploring the correlation of features in a dataset.

To get the correlation of the features inside a dataset we can call `<dataset>.corr()` , which is a Pandas dataframe method. This will give use the correlation matrix.

We can now use either Matplotlib or Seaborn to create the heatmap.

Matplotlib:

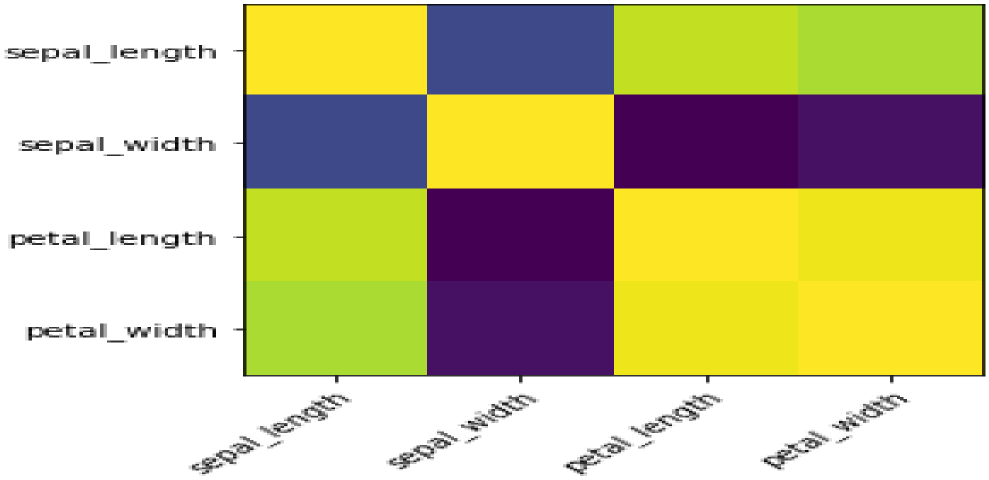


Figure 20 - Matplotlib

7.Heatmap without annotations

Data visualization is the discipline of trying to understand data by placing it in a visual context, so that patterns, trends and correlations that might not otherwise be detected can be exposed.Python offers multiple great graphing libraries that come packed with lots of different features. In this article we looked at Matplotlib, Pandas visualization and Seaborn.

2.3 ANACONDA NAVIGATOR

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda distribution that allows you to launch applications and easily manage conda packages, environments and channels without using command-line commands. Navigator can search for packages on Anaconda Cloud or in a local Anaconda Repository. It is available for Windows, mac OS and Linux.

2.3.1 Why use Navigator?

In order to run, many scientific packages depend on specific versions of other packages. Data scientists often use multiple versions of many packages, and use multiple environments to separate these different versions.

The command line program conda is both a package manager and an environment manager, to help data scientists ensure that each version of each package has all the dependencies it requires and works correctly.

Navigator is an easy, point-and-click way to work with packages and environments without needing to type conda commands in a terminal window. You can use it to find the packages you want, install them in an environment, run the packages and update them, all inside Navigator.

2.3.2 what applications can i access using navigator?

The following applications are available by default in Navigator:

- Jupyter Lab
- Jupyter Notebook
- QT Console
- Spyder
- VS Code
- Glue viz
- Orange 3 App
- Rodeo
- RStudio

Advanced conda users can also build your own Navigator applications

2.3.3 How can I run code with Navigator?

The simplest way is with Spyder. From the Navigator Home tab, click Spyder, and write and execute your code.

You can also use Jupyter Notebooks the same way. Jupyter Notebooks are an increasingly popular system that combine your code, descriptive text, output, images and interactive interfaces into a single notebook file that is edited, viewed and used in a web browser

CHAPTER-3

SYSTEM ANALYSIS

3.1 EXISTING SYSTEM

Heart diseases have emerged as one of the most prominent cause of death all around the world. According to World Health Organisation , heart related diseases are responsible for the taking 17.7 million lives every year, 31% of all global deaths. In India too, heart related diseases have become the leading cause of mortality [1]. Heart diseases have killed 1.7 million Indians in 2016, according to the 2016 Global Burden of Disease Report, released on September 15,2017. Heart related diseases increase the spending on health care and also reduce the productivity of an individual. Estimates made by the World Health Organisation (WHO), suggest that India have lost up to $237 billion, from 2005-2015, due to heart related or Cardiovascular diseases.Thus, feasible and accurate prediction of heart related diseases is very important .Medical organisations, all around the world, collect data on various health related issues. These data can be exploited using various machine learning techniques to gain useful insights. But the data collected is very massive and, many a times, this data can be very noisy. These datasets, which are too overwhelming for human minds to comprehend, can be easily explored using various machine learning techniques. Thus, these algorithms have become very useful, in recent times, to predict the presence or absence of heart related diseases accurately.

3.2 PROPOSED SYSTEM

Dimensionality Reduction involves selecting a mathematical representation such that one can relate the majority of, but not all, the variance within the given data, thereby including only most significant information. The data considered for a task or a problem, may consists of a lot of attributesor dimensions, but not all of these attributes may equally influence the output. A large number of attributes, or features, may affect the computational complexity and may even lead to overfitting which leads to poor results. Thus, Dimensionality Reduction is a very important step considered while building any model.

3.2.1 Advantages

- User can search for doctor's help at any point of time.
- User can talk about their Heart Disease and get instant diagnosis.
- Doctors get more clients online.
- Very useful in case of emergency.

3.3 SYSTEM ARCHITECTURE

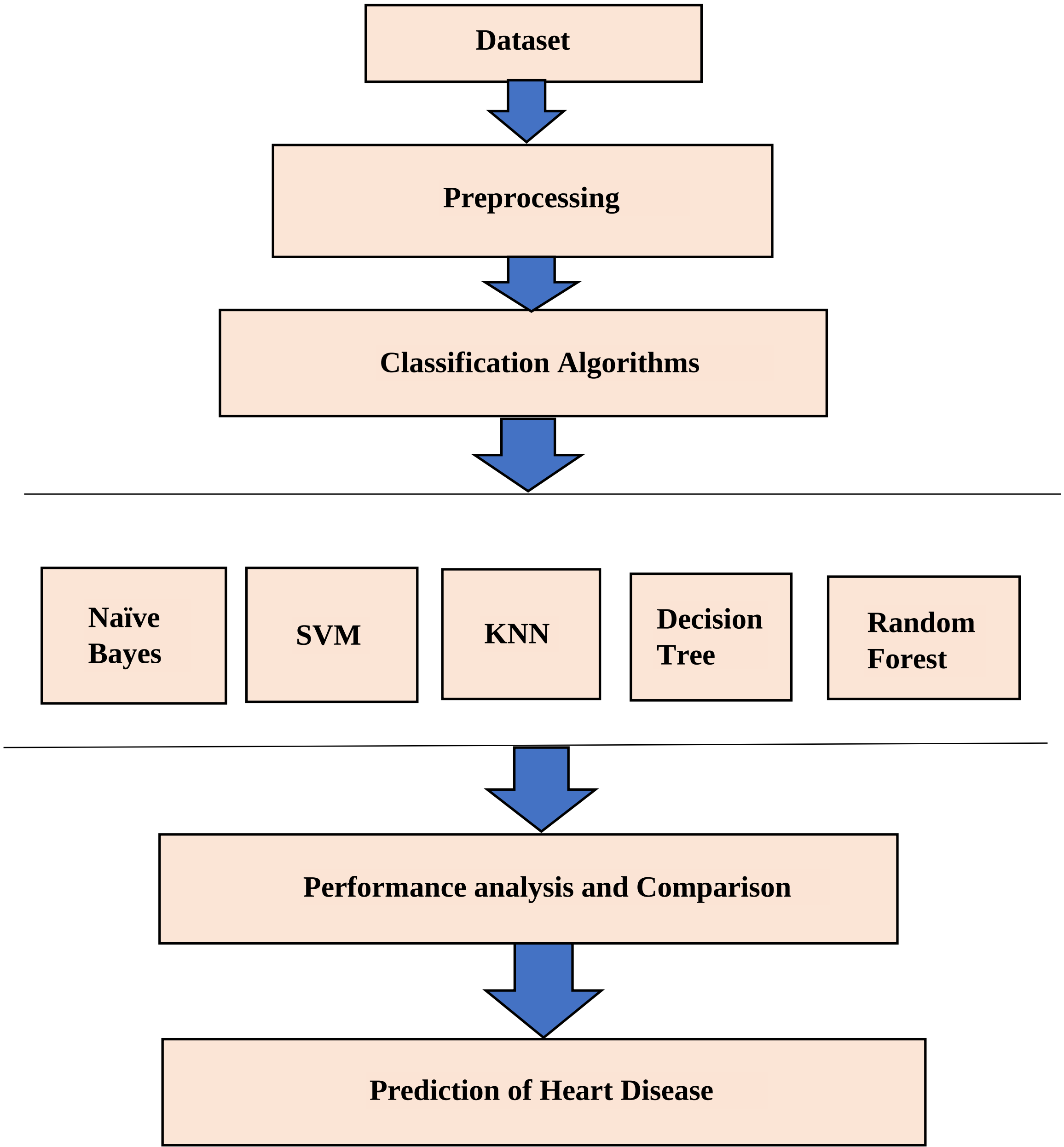

Dataset
Preprocessing
Classification Algorithms
Naïve Bayes
SVM
KNN
Decision Tree
Random Forest
Performance analysis and Comparison
Prediction of Heart Disease

Figure 21-System Architecture

3.4 FEASIBILITY STUDY

The idea of this research work is to study the various prediction models for heart disease and improve the performance of the model by using genetic algorithm. The performances of the models are tested with four region heart disease data sets such as Cleavelan d, Hungarian, Lonf-beach and Switzerland datasets. The dataset used in this work is obtained from UCI machine learning repository Heart Disease database contains attributes, but mostly all published experiments refer to using a subset for the prediction. The proposed work considered risk factors hypertension and family history as predictor Risk factors such as smoking, obesity, diabetes are having missing values in most of the data

3.5 REQUIREMENT SPECIFICATION

3.5.1 Functional Requirements

▪ Graphical User interface with the User.

3.5.2 Software Requirements

1. Python

2. Anaconda Navigator

3. Jupiter notebook

3.5.2.1 Operating Systems supported

Windows 7

3.5.2.2 Debugger and Emulator

▪ Any Browser (Particular Chrome)

3.5.3 Hardware Requirement

▪ Processor: Pentium IV or higher

▪ RAM: 4GB

▪ Space on Hard Disk: minimum 512M.

CHAPTER- 4

SYSTEM DESIGN

4.1 UML Diagrams

4.1.2 OBJECT DIAGRAM

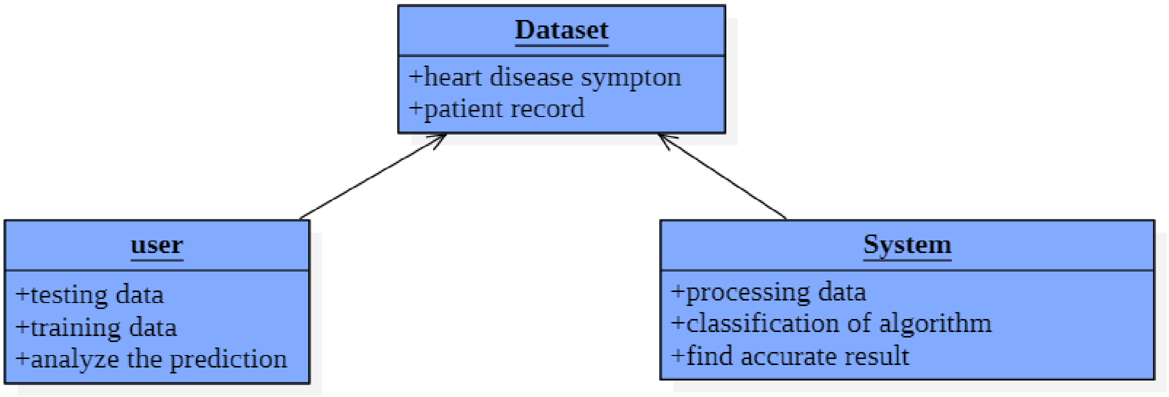


Figure 22-Object Diagram

The figure represents the object diagram. An object diagram shows a complete or partial view of the structure of a modeled system at a specific time. Here objects are user, Dataset, system. User and system do the processing on dataset.

4.1.2 CLASS DIAGRAM

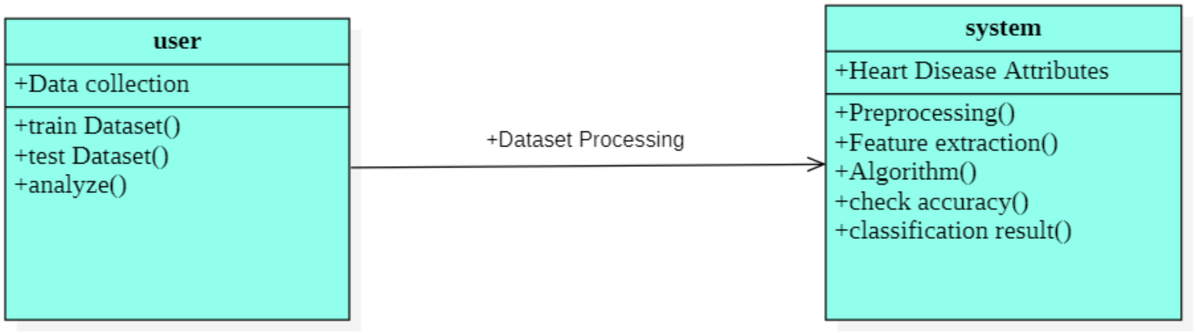


Figure 23-Class Diagram

The figure represents class diagram. A class diagram is static structure diagram that specifies the structure of system by showing the systems classes ,their attributes, operations and relationships among objects. Here user, system are the classes ,their attributes and operations are shown in the figure.

4.1.3 USECASE DIAGRAM

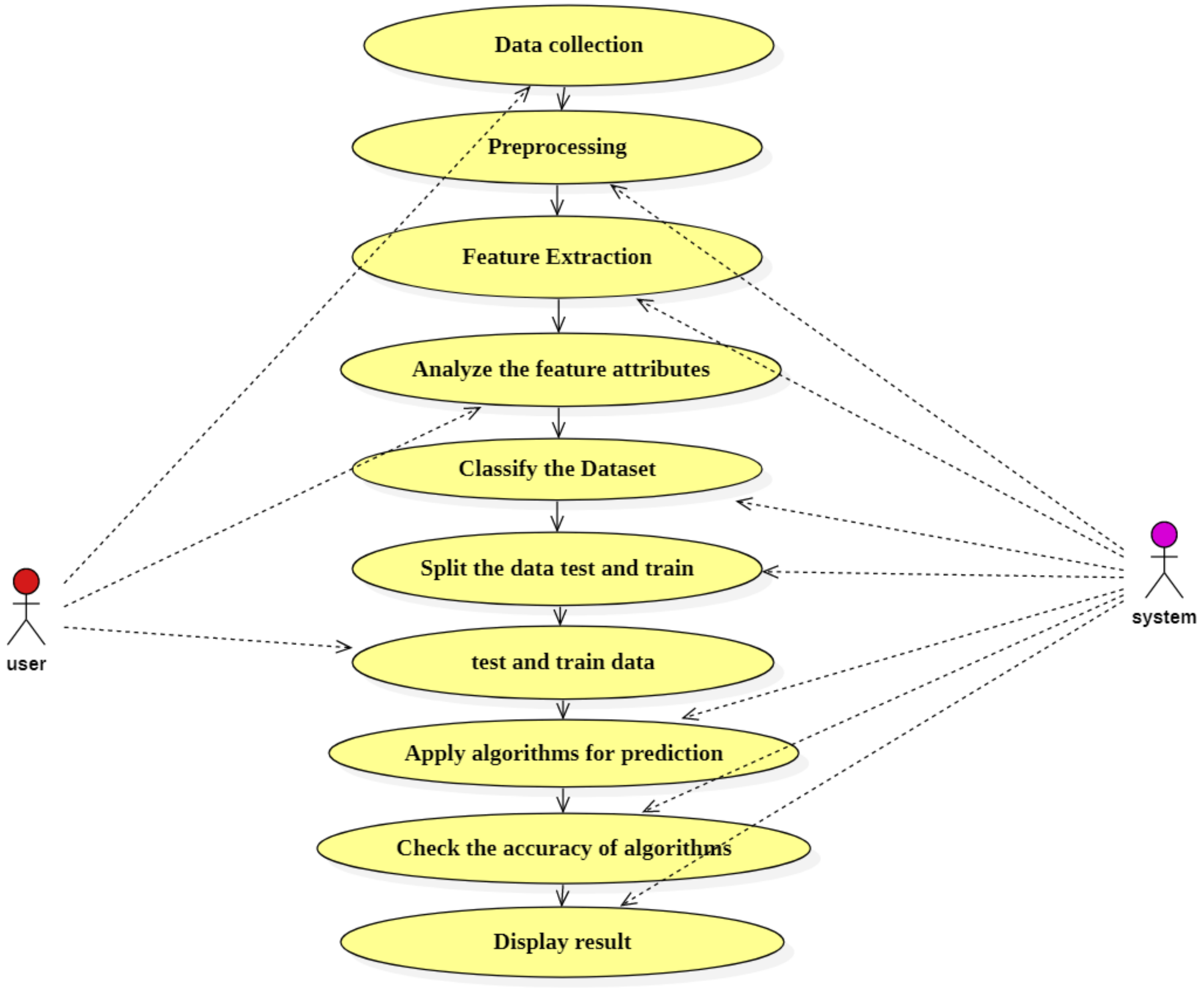

Data collection
Preprocessing
Feature Extraction
Analyze the feature attributes
Classify the Dataset
Split the data test and train
test and train data
Apply algorithms for prediction
Check the accuracy of algorithms
Display result
user
system

Figure 24 -Usecase Diagram

The figure shows the Usecase diagram. Here the actors are user and system. Usecases represents the actions performed by the both user and system. Actor as a user can collect the data of heart disease of patients record. Actor as system can do preprocessing and extracting the feature attributes of dataset. User analyze the feature attributes , system split the data to test and train. Testing and Training done by the user.Apply algorithms for prediction. System provide the accuracy score of algorithms and display result to the user.

4.1.4 SEQUENCE DIAGRAM

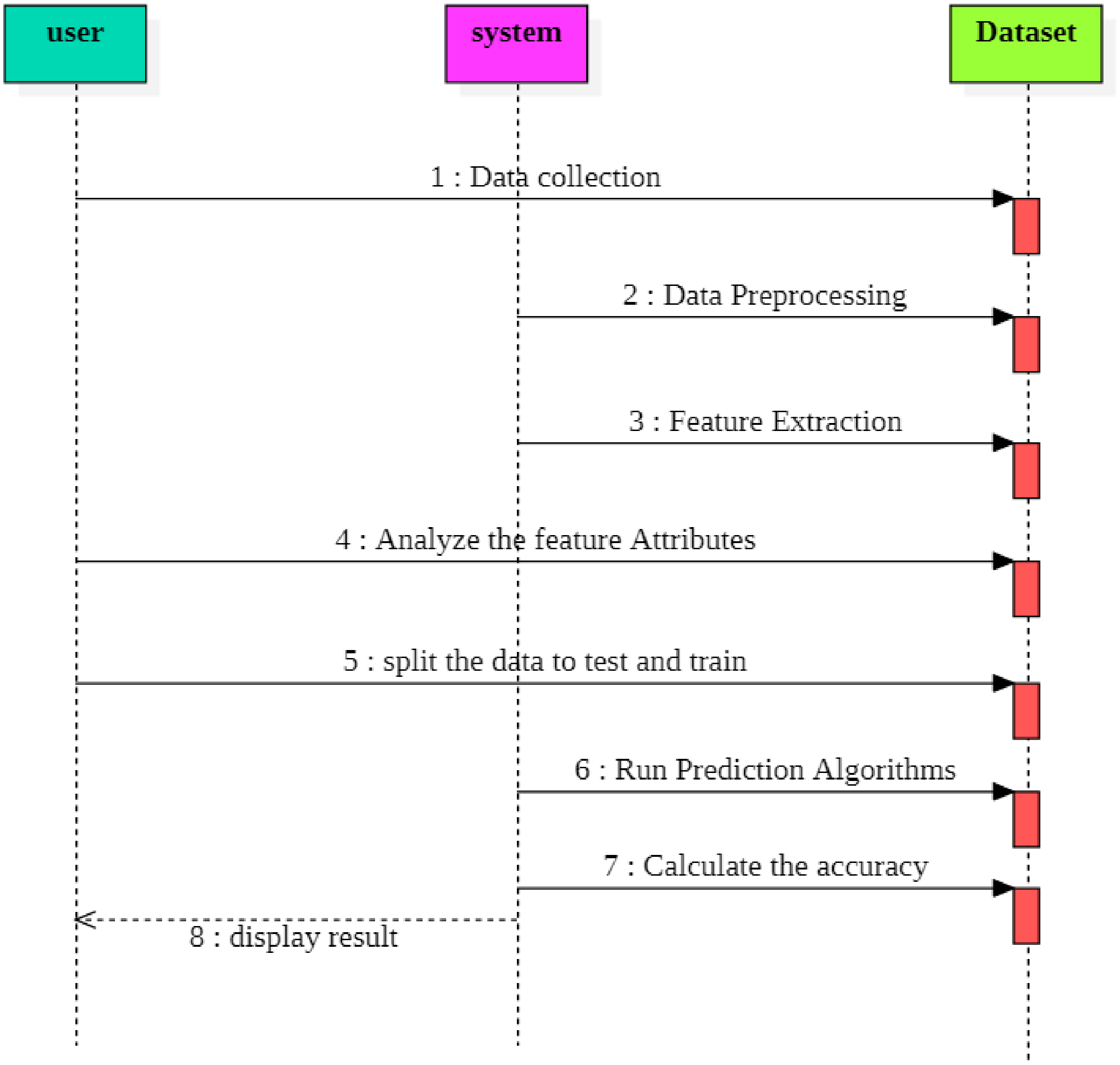

user
system
Dataset
1 : Data collection
2 : Data Preprocessing
3 : Feature Extraction
4 : Analyze the feature Attributes
5 : split the data to test and train
6 : Run Prediction Algorithms
7 : Calculate the accuracy
8 : display result

Figure 25 -Sequence Diagram

The figure represents the Sequence diagram. It shows the object interaction arranged in time sequence. The parallel vertical line are lifelines, different objects that live simultaneously and horizontal lines are messages. The sequence of messages exchanged among object. Here user, system and Dataset are the objects. The messages are exchanged between them. User can collect the data, analyze the feature attributes, split the data to test and train from dataset by passing messages. System can preprocess the data, extracting the features from dataset, applying algorithms for prediction ,calculate the accuracy of all algorithms and display result to the user

4.1.5 COMMUNICATION DIAGRAM

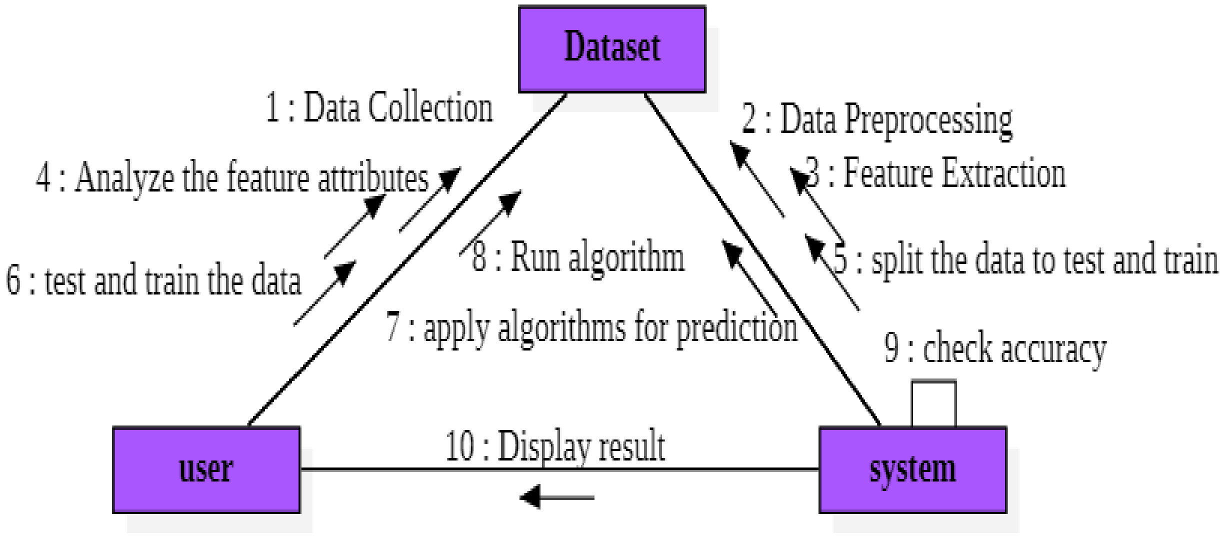


Figure 26 -Communication Diagram

The above figure represents the communication diagram. Communication diagrams, like the sequence diagrams - a kind of interaction diagram, shows how objects interact. A communication diagram is an extension of object diagram that shows the objects along with the messages that travel from one to another. In addition to the associations among objects, communication diagram shows the messages the objects send each other. Here the communication between the user, system and dataset by passing messages to each other.

4.1.6 ACTIVITY DIAGRAM

The figure represents the Activity diagram. It is a behavioural diagram . It depicts behaviour of the system. An activity diagram portrays the control flow from a start point to a finish point showing the various decision paths that exist while the activity is being executed. Here the activity start with the importing the libraries then importing dataset of Heart Disease patient after that data preprocessing and feature extraction, analyzing the feature attributes for prediction, split the dataset to test and train, apply algorithms ,run the algorithms and check the accuracy of all algorithms ,display result to user.

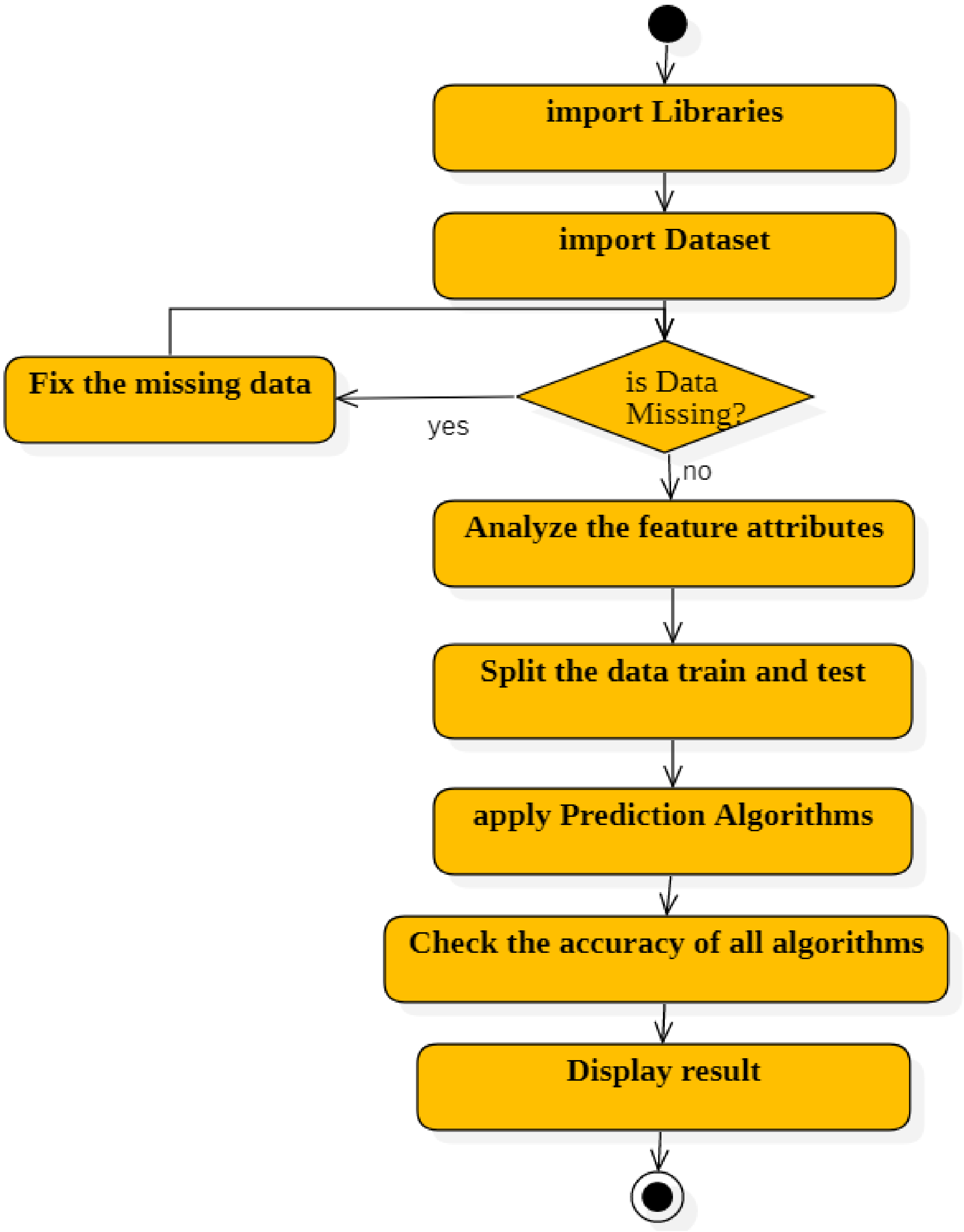

import Libraries
import Dataset
Fix the missing data
is Data Missing?
yes
no
Analyze the feature attributes
Split the data train and test
apply Prediction Algorithms
Check the accuracy of all algorithms
Display result

Figure 27 -Activity Diagram

4.1.7 STATECHAERT DIAGRAM

The figure represents the State chart diagram. A Sate chart diagram is used to represent the condition of the system or part of the system at the finite instances of time. It is a behavioural diagram and it represents the behaviour using finite state transitions. It specifies the sequences of states that an object can be in event and condition. Here the user behaviour can be represented as state transitions. The user perform action can be represented in sequence of states. The states are the action done by the user such as data collection ,analyze the feature attributes, test and train the data. The states of action done by the system are data preprocessing, Feature extraction ,apply algorithms for prediction, calculate the accuracy of algorithm and display the result to user.

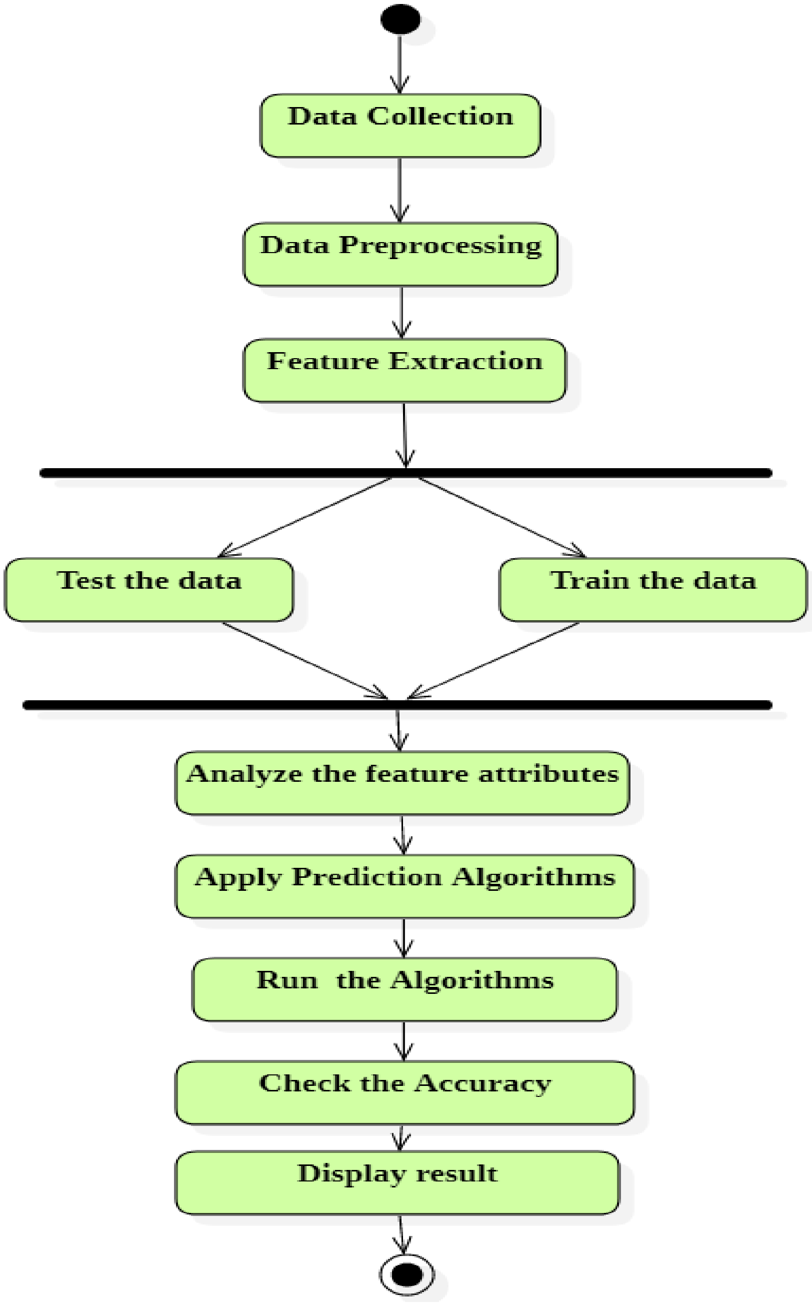


Figure 28 -Statechart Diagram

4.1.8 COMPONENT DIAGRAM

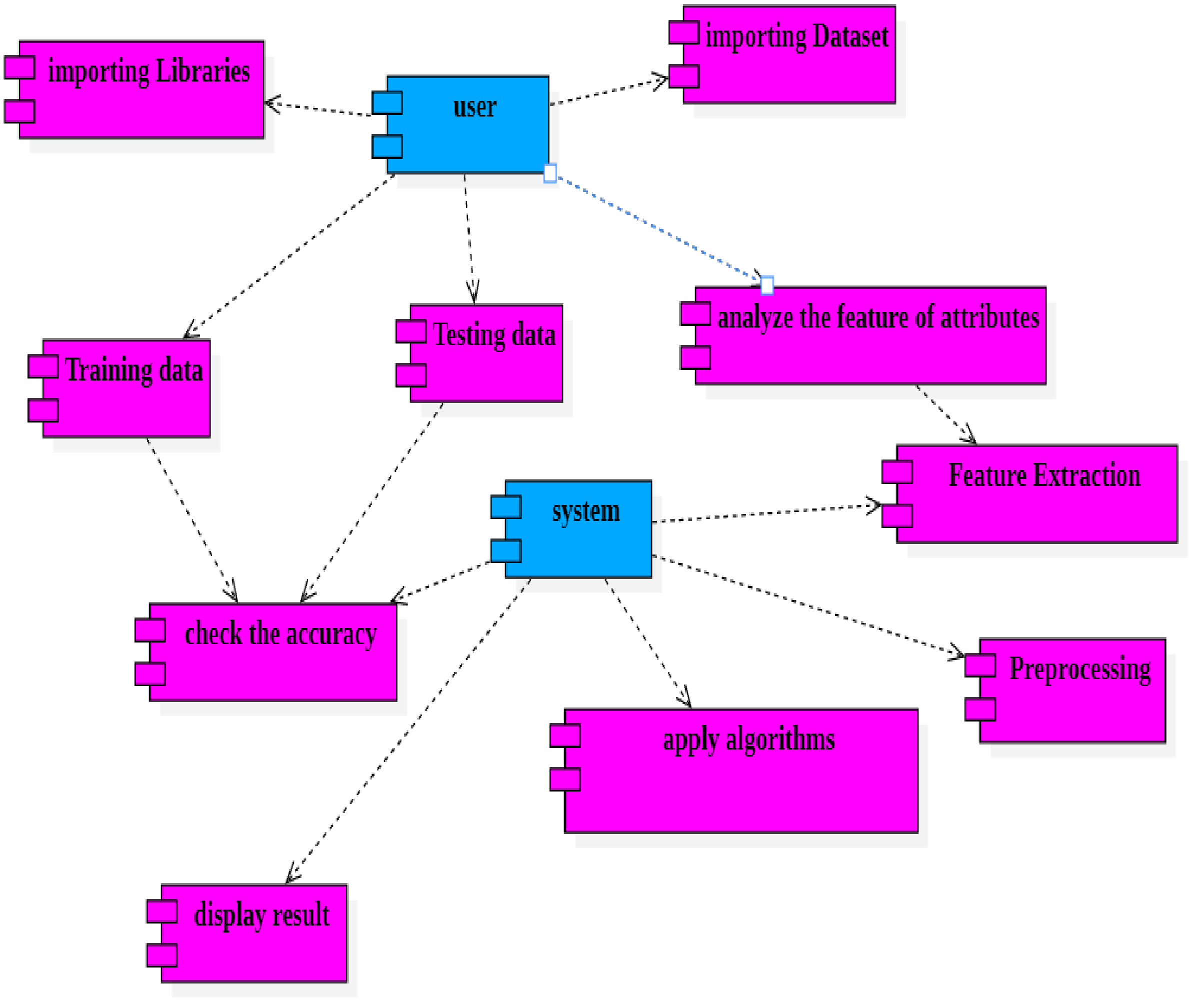


Figure 29- Component Diagram

The figure represents the component diagram. Component diagram are used for modelling subsystems. Its represents how each and every components acts during execution and running of the system program. They are also used to show and represent structure and organization of all components. Here the components are user and system, importing libraries, importing dataset, Training data and Testing data ,preprocessing , analyze the feature attributes ,apply algorithms ,check the accuracy ,display results.

4.1.9 DEPLOYMENT DIAGRAM

The figure represents the Deployment diagram. The deployment diagram visualizes the physical hardware on which the software will be deployed. It involves the nodes and their relationship. It maps the software architecture, where the software will be executed as node. since it involves many nodes and their relations is shown by utilizing communication path. Here the nodes are user and system are doing processing on the dataset.

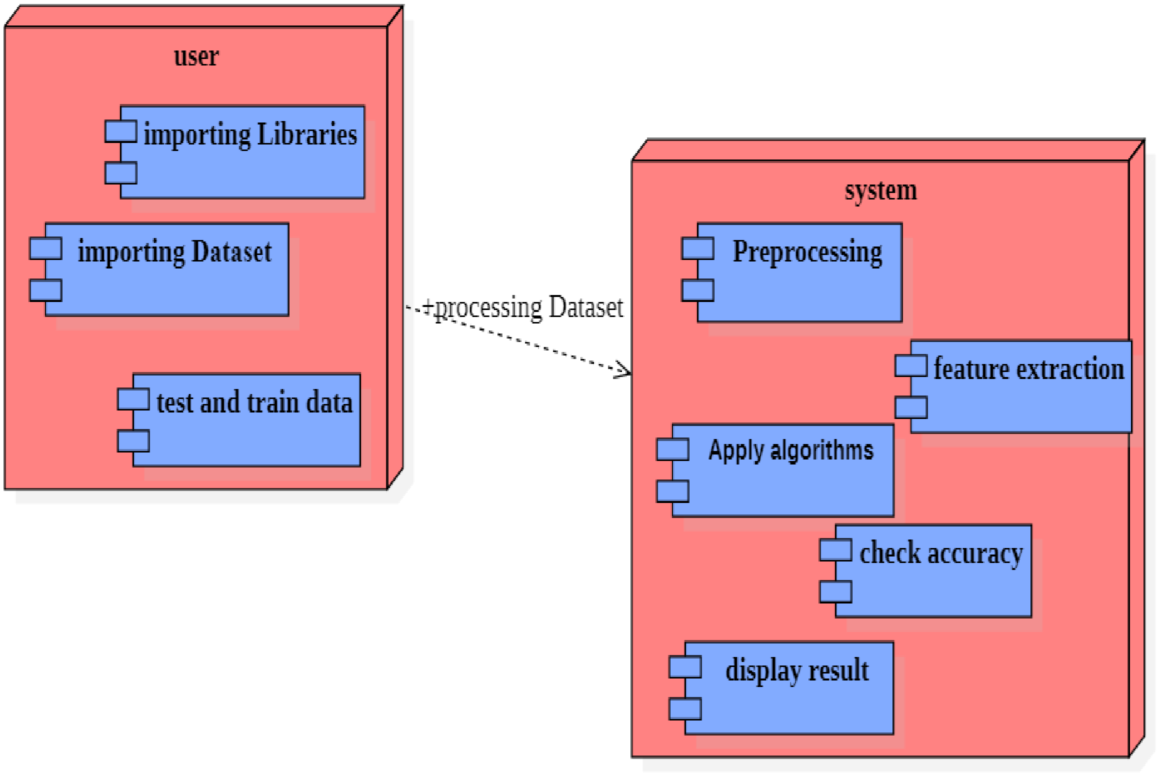


Figure 30-Deployment Diagram

CHAPTER-5

IMPLEMENTATION AND CODING

5.1 MODULES

- User
- System
- Dataset

5.2 MODULES DESCRIPTION

1.User Description- User can collect the dataset which contain Heart Disease Patients record with different Heart Disease Symptoms and contain various attributes for prediction.

2.System Description- System can do data preprocessing, feature extraction, apply algorithm for prediction ,calculate the accuracy score of all algorithms and display the result to the user.

3.Dataset Description- Our dataset is based on UCI heart Disease Data Set [6] and we have 303 instances. According to UCI, "This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. We guess too many features will bring too much noise so people has done feature extraction and reduce 76 features to 14 features. To better understand the meaning of the features, we have the responsibility to explain some of the attributes of original dataset from UCI as follows:

• age: age in years

• sex: sex (1 = male; 0 = female)

• cp: chest pain type -- Value 0: typical angina -- Value 1: atypical angina -- Value 2: non-anginal pain -- Value 3: asymptomatic

• trestbps: resting blood pressure (in mm Hg on admission to the hospital)

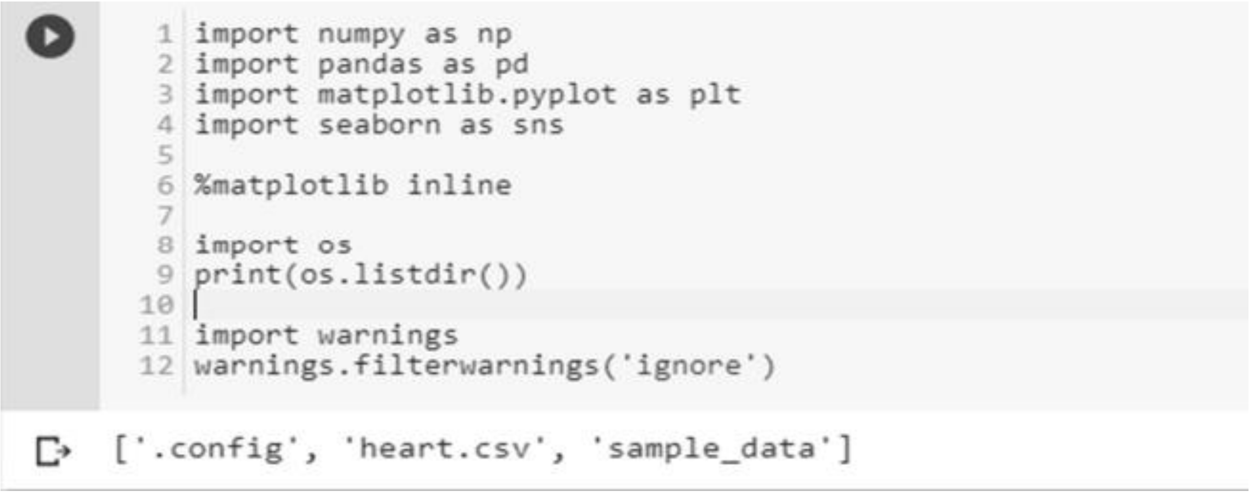
• chol: serum cholestoral in mg/dl

• fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

• target: Heart disease (0 = no, 1 = yes)

Since the original dataset has missing values, we just downloaded a clean dataset from Kaggle[7]. We have split the dataset into 80% (242 instances) for training and 20%(61 instances) for test. We did normalization on our dataset to avoid overfitting. What we did to our dataset is to change 1s to 0s in target column and vice versa in order to make value 1 indicate the presence of heart disease and make value 0 indicate the absence of heart disease. Given such dataset we can do many interesting predicative tasks. For example, we can use these features to predict chest pain type. But the most important thing is that given the 13 attributes from a patient, we want to predict whether he has the heart disease or not because keeping healthy is very import to people.

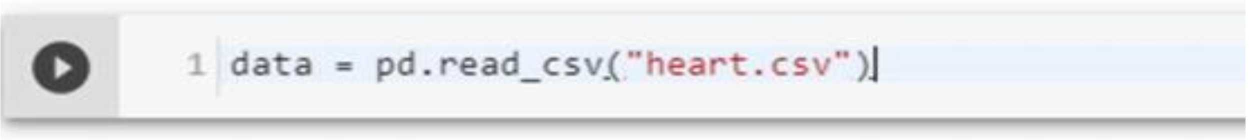
3.1 Importing libraries



```
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')
```

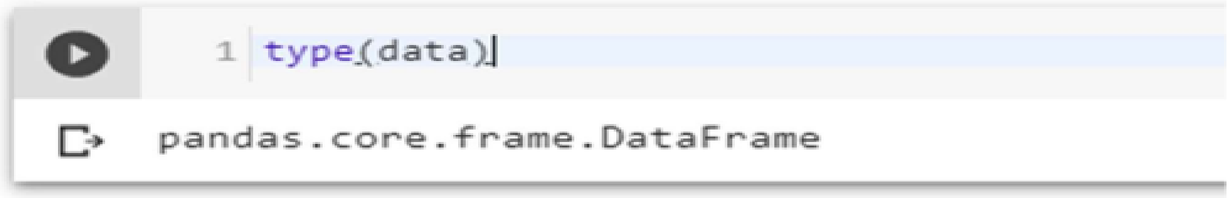
```
['.config', 'heart.csv', 'sample_data']
```

3.2 Load dataset



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

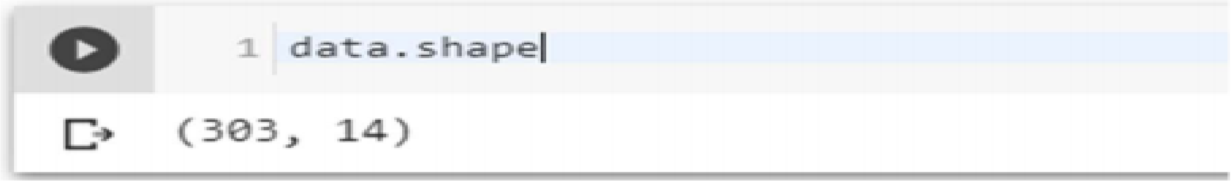
3.3 Check the type of dataset



```
type(data)
```

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

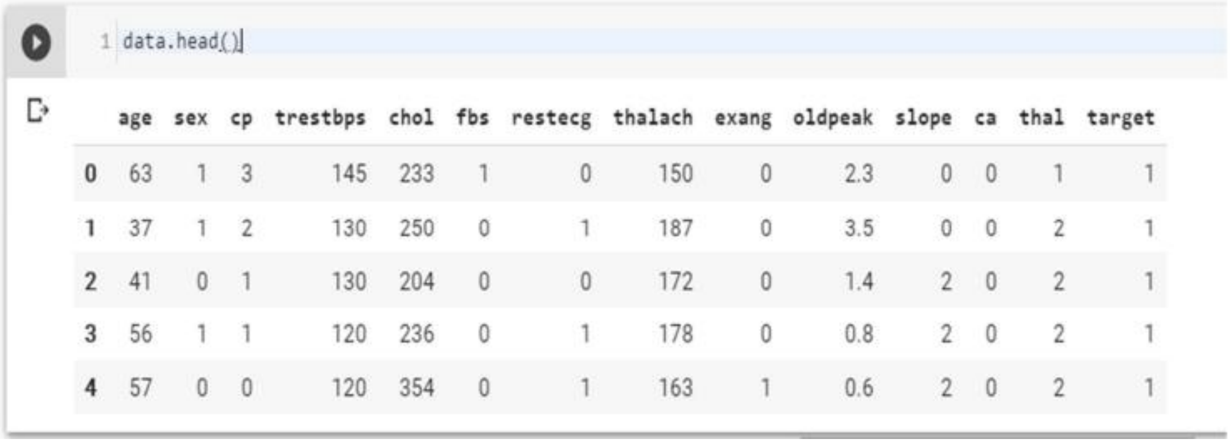
3.4 Check the shape of dataset



```
data.shape
```

```
(303, 14)
```

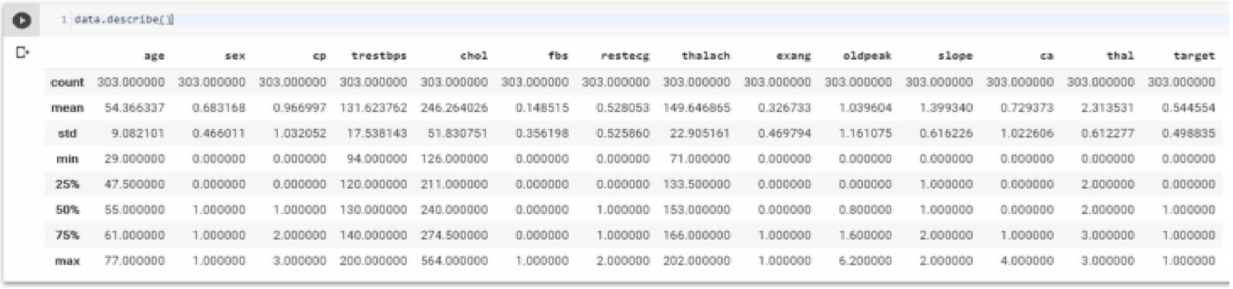
3.5 Check the 5 columns of dataset



```
data.head()
```

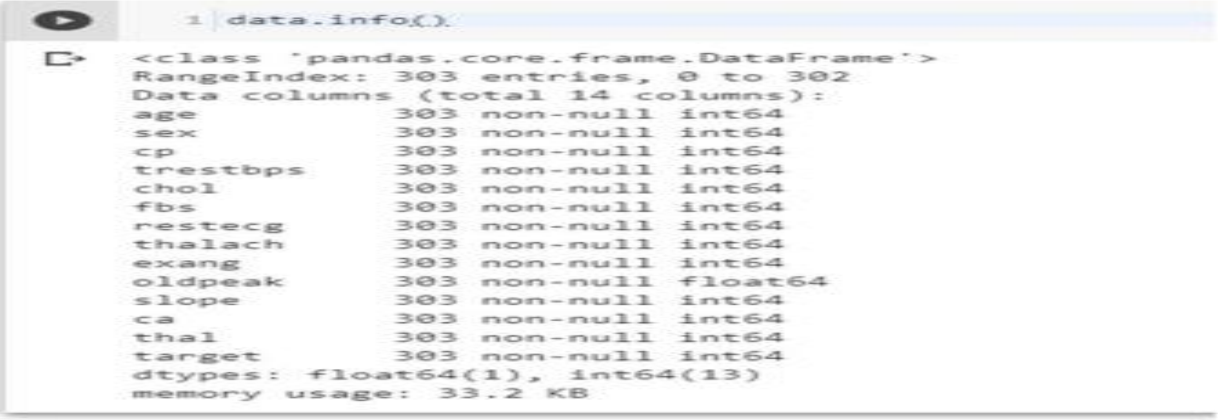
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

3.6 Dataset Description



```
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



```
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 = nonanginal 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)

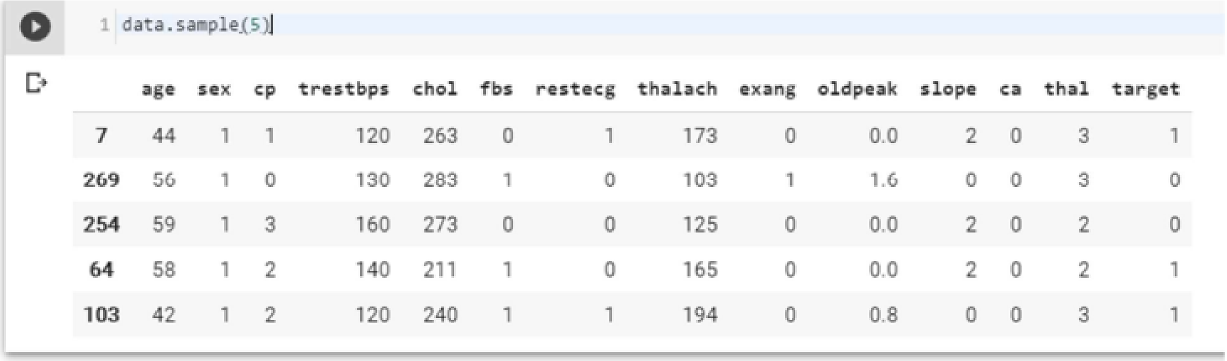
3.4.7 Types of features

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

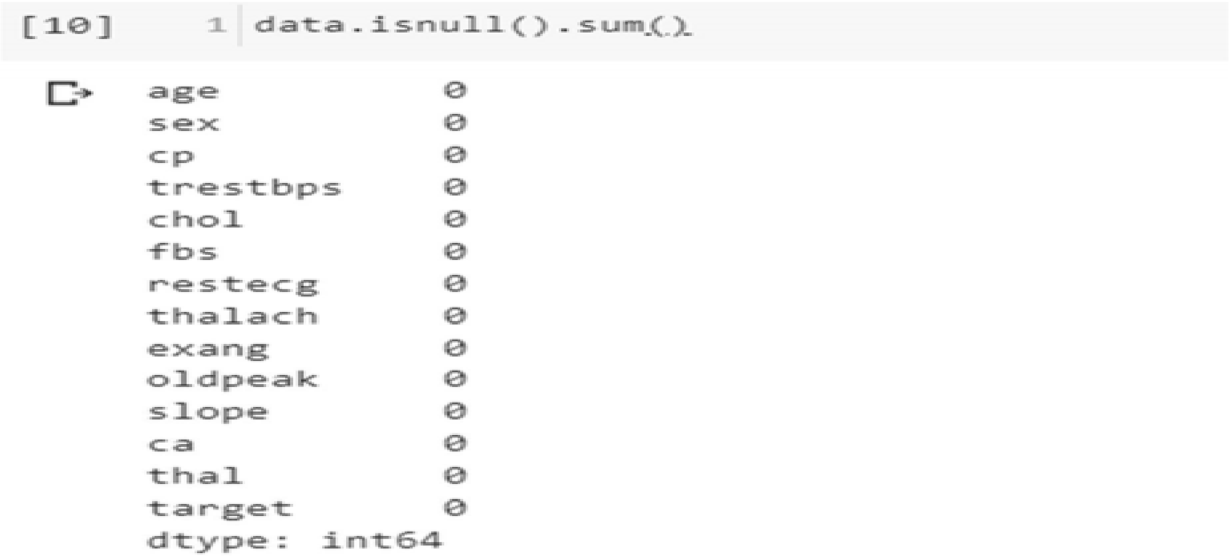


```
data.sample(5)
```

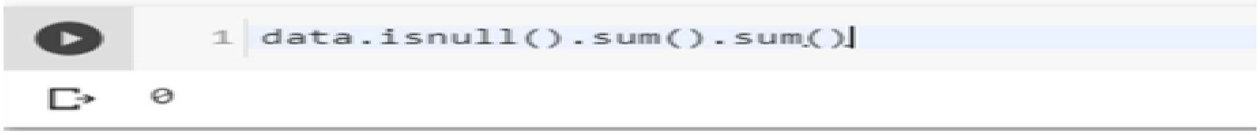
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
7	44	1	1	120	263	0	1	173	0	0.0	2	0	3	1
269	56	1	0	130	283	1	0	103	1	1.6	0	0	3	0
254	59	1	3	160	273	0	0	125	0	0.0	2	0	2	0
64	58	1	2	140	211	1	0	165	0	0.0	2	0	2	1
103	42	1	2	120	240	1	1	194	0	0.8	0	0	3	1

Data Preprocessing

3.4.9 Check for missing Data



```
[10]    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
```

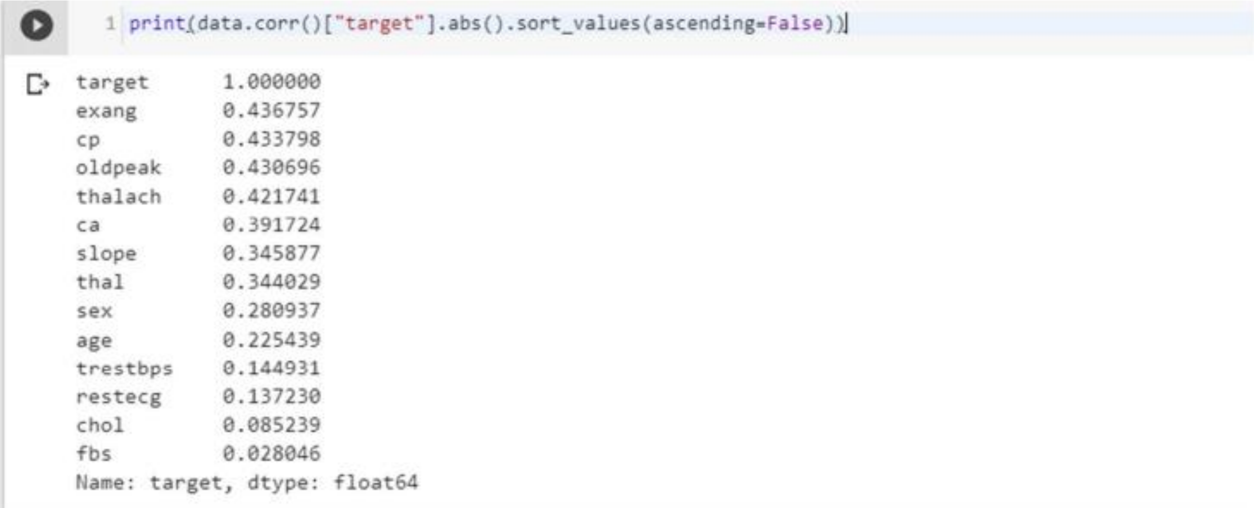


```
data.isnull().sum().sum()
```

```
0
```

No Data is missing, which is good.

3.10 Check the correlation with target data

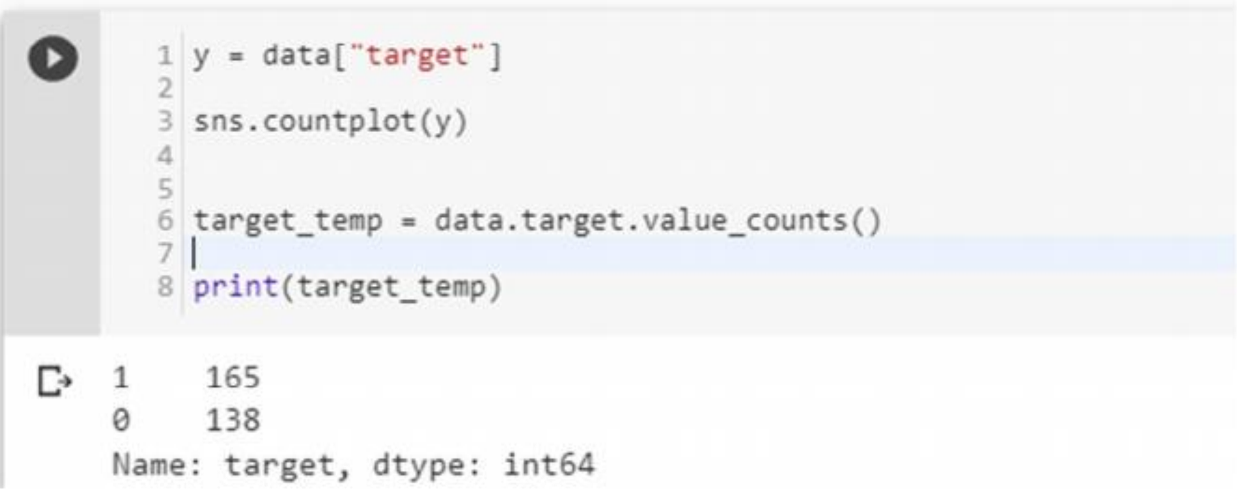


```
print(data.corr()["target"].abs().sort_values(ascending=False))
```

```
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 show that most columns are moderately correlated with target ,but ‘fbs’ is very weakly correlated

3.11 Exploratory Data Analysis (EDA)

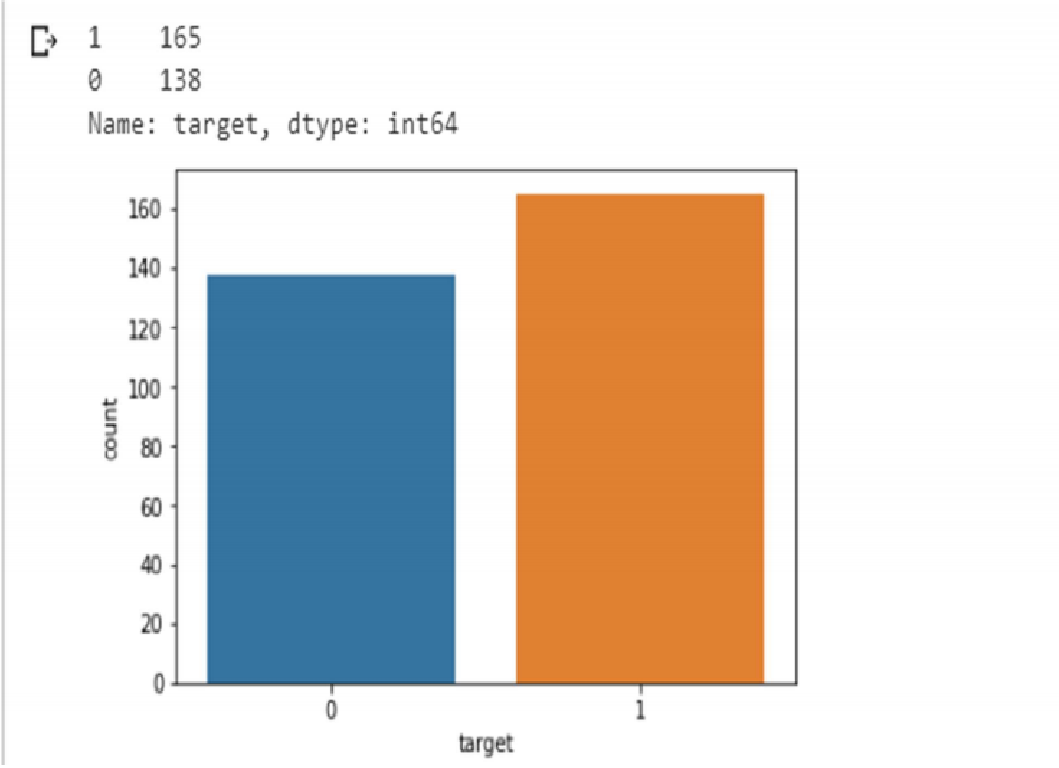


```
y = data["target"]

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)

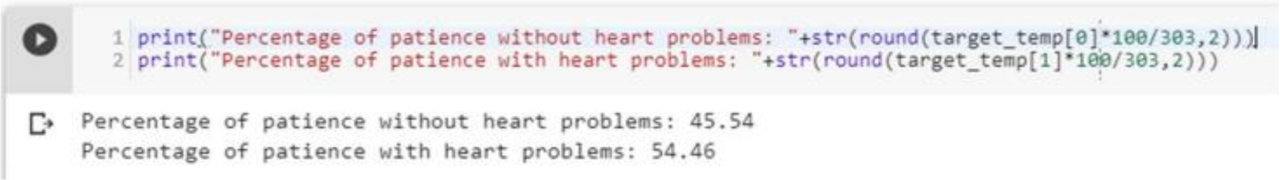
No. of Heart Disease patients is 165. No. of patients who don’t have a heart disease is 138. [Which is a good balance of target data.]



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

Figure 31 Disease vs Non-Disease Analysis

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

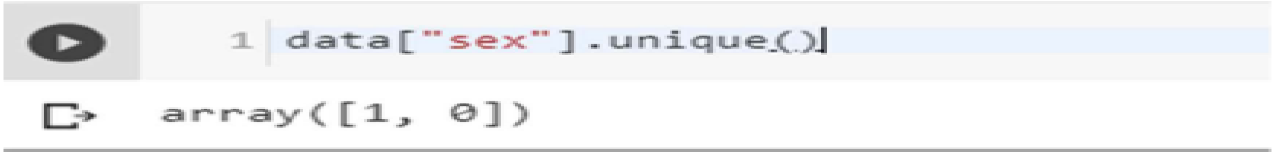


```
print("Percentage of patience without heart problems: "+str(round(target_temp[0]*100/303,2)))
print("Percentage of patience with heart problems: "+str(round(target_temp[1]*100/303,2)))
```

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

3.11.2 Uniqueness of sex column

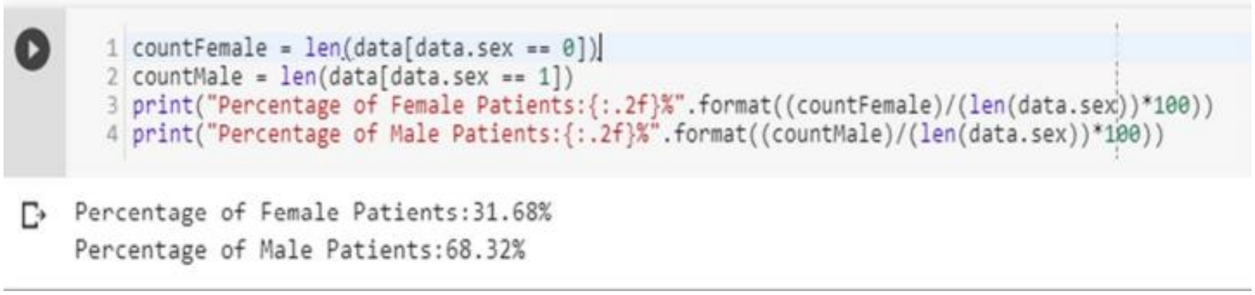
– Two sex types: 1 is male and 0 is female



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

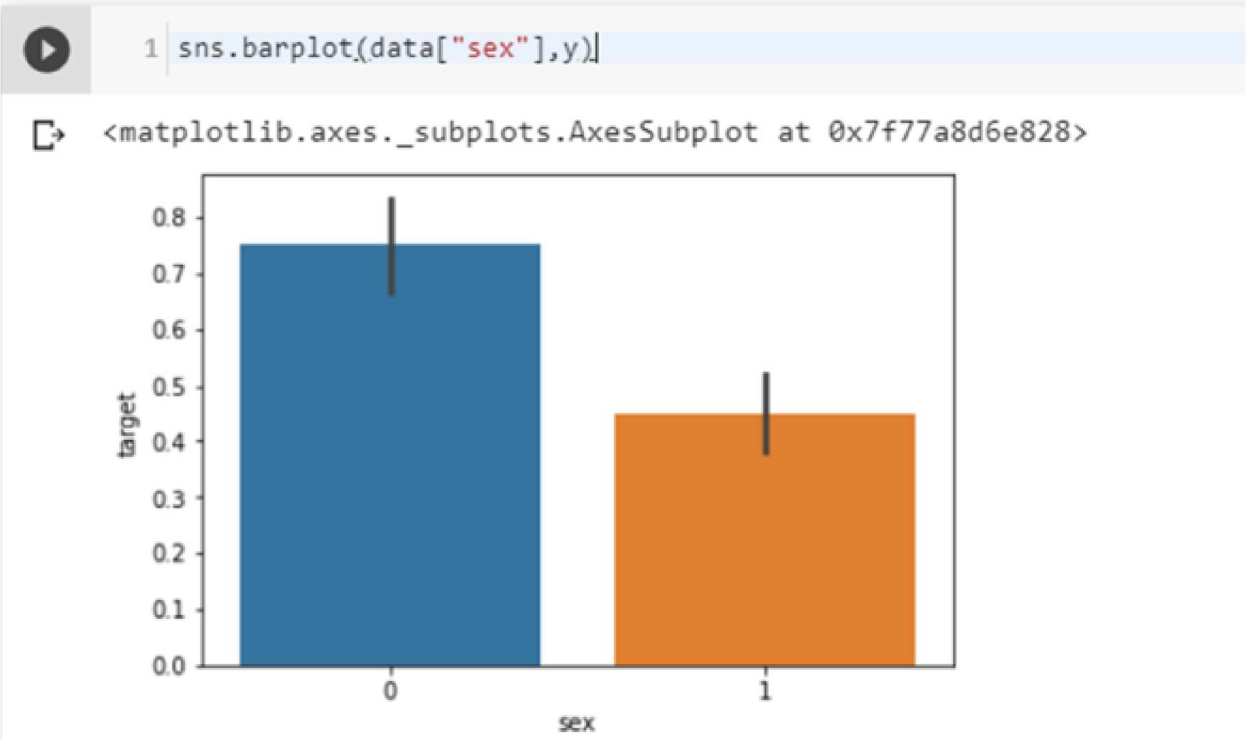
```
array([1, 0])
```

3.11.3 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%
```

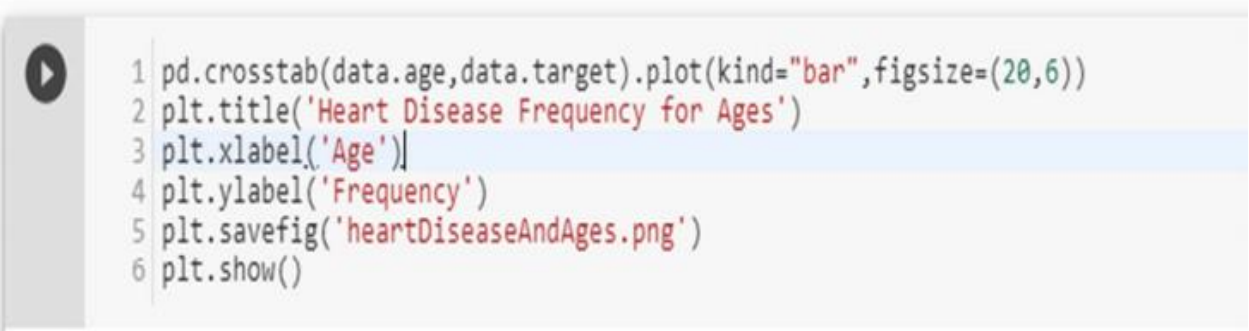


```
sns.barplot(data["sex"],y)
```

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

Figure 32- Sex analysis

3.11.4 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()
```

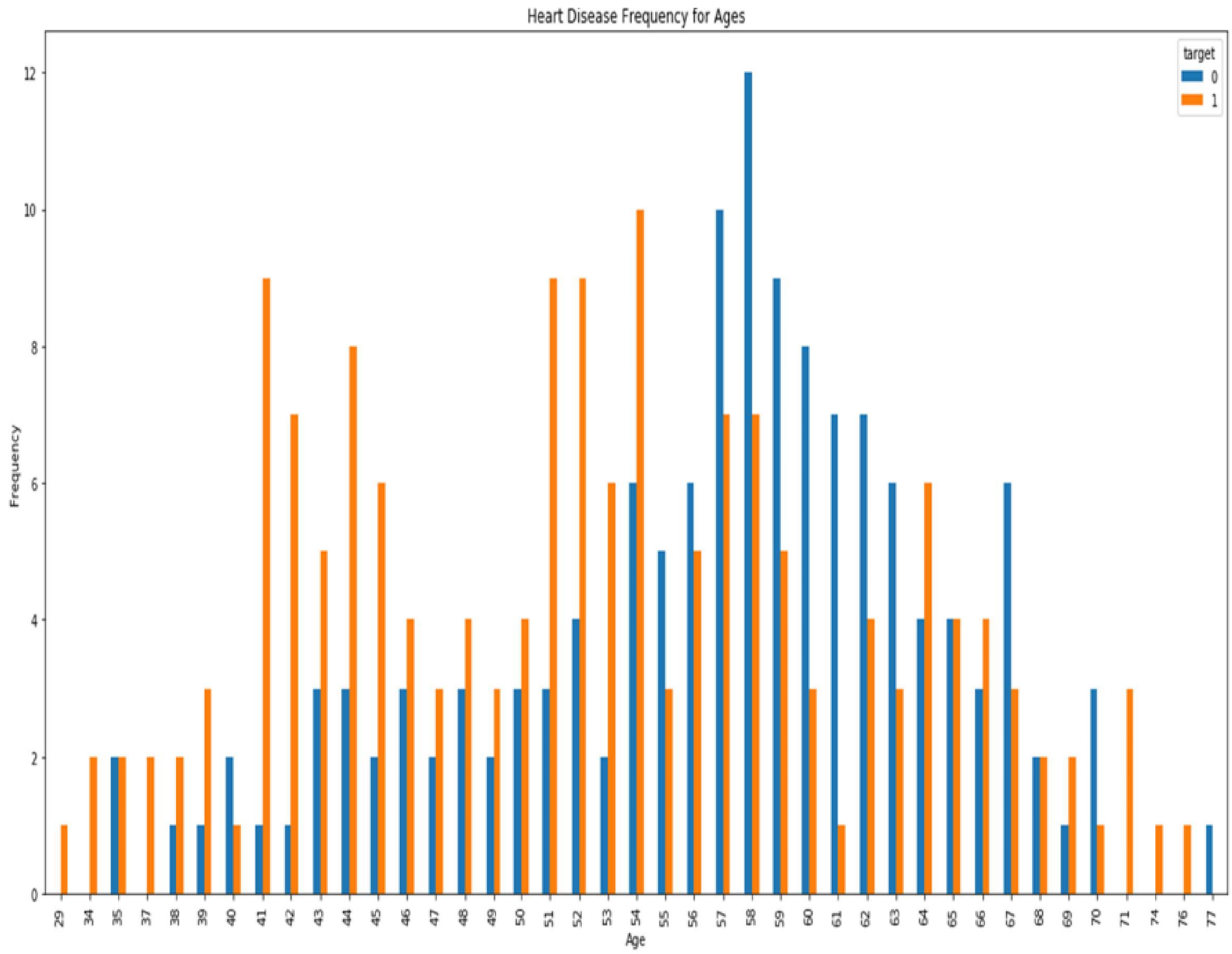
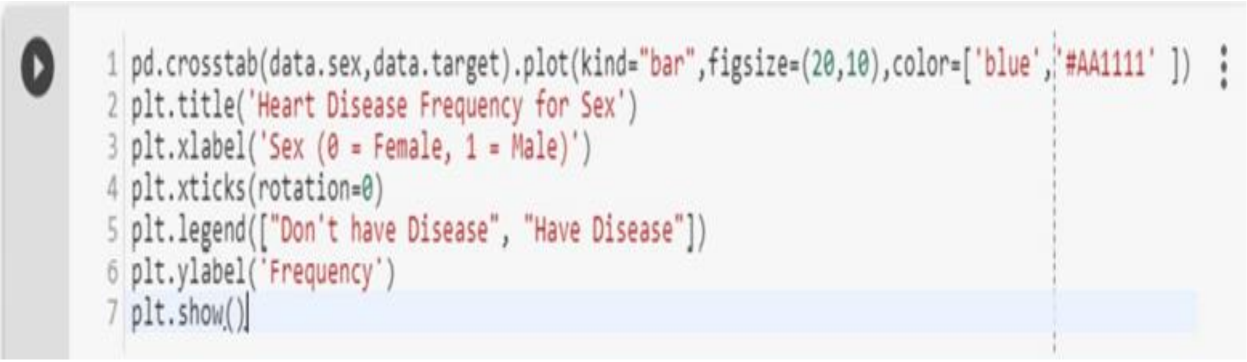


Figure 33 - Heart Disease Frequency for ages

3.11.5 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”

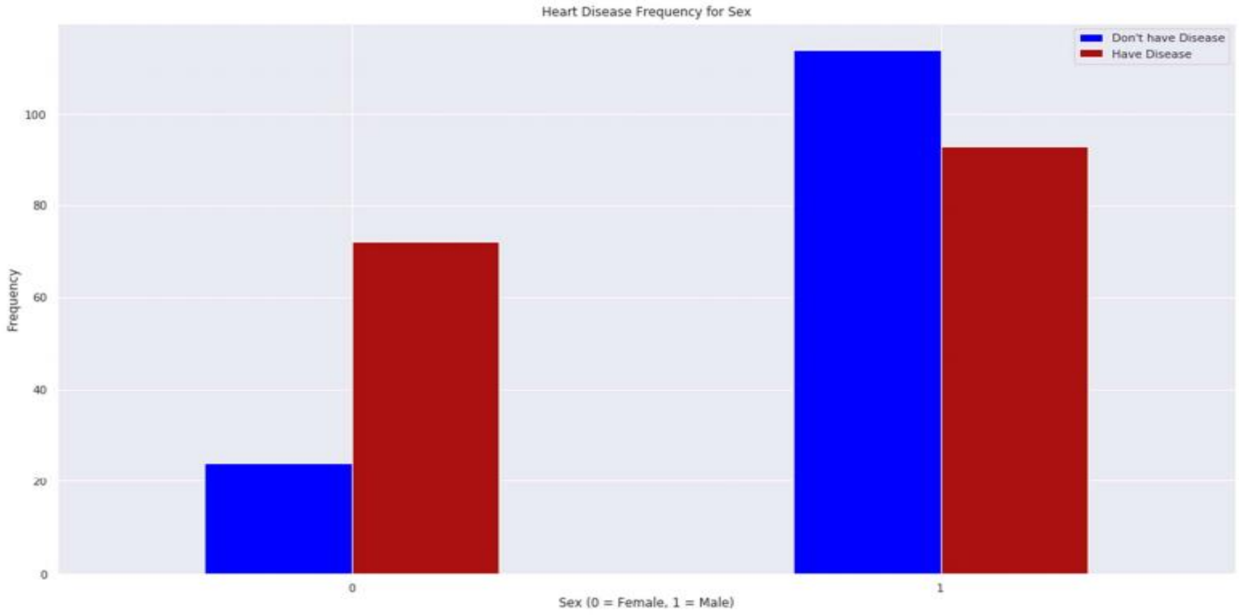
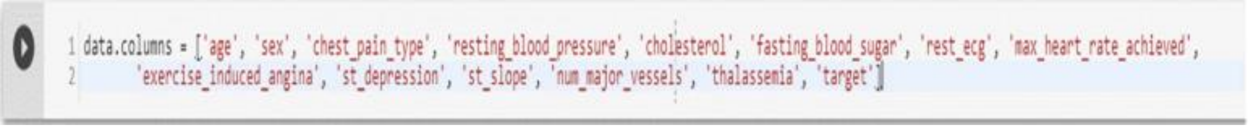


Figure 34 - Heart Disease Frequency for sex

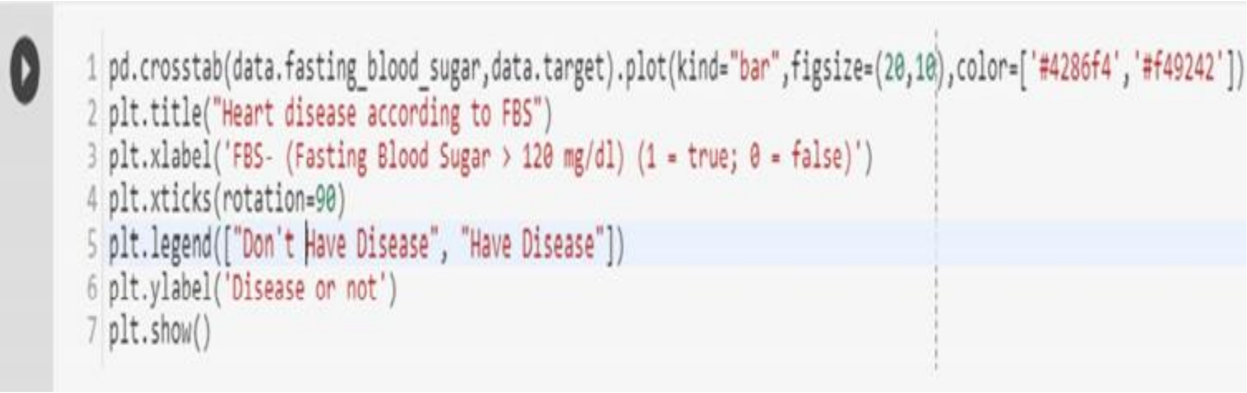
3.11.6 Making the data column names easily recognizable



```
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']
```

3.11.7 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()
```

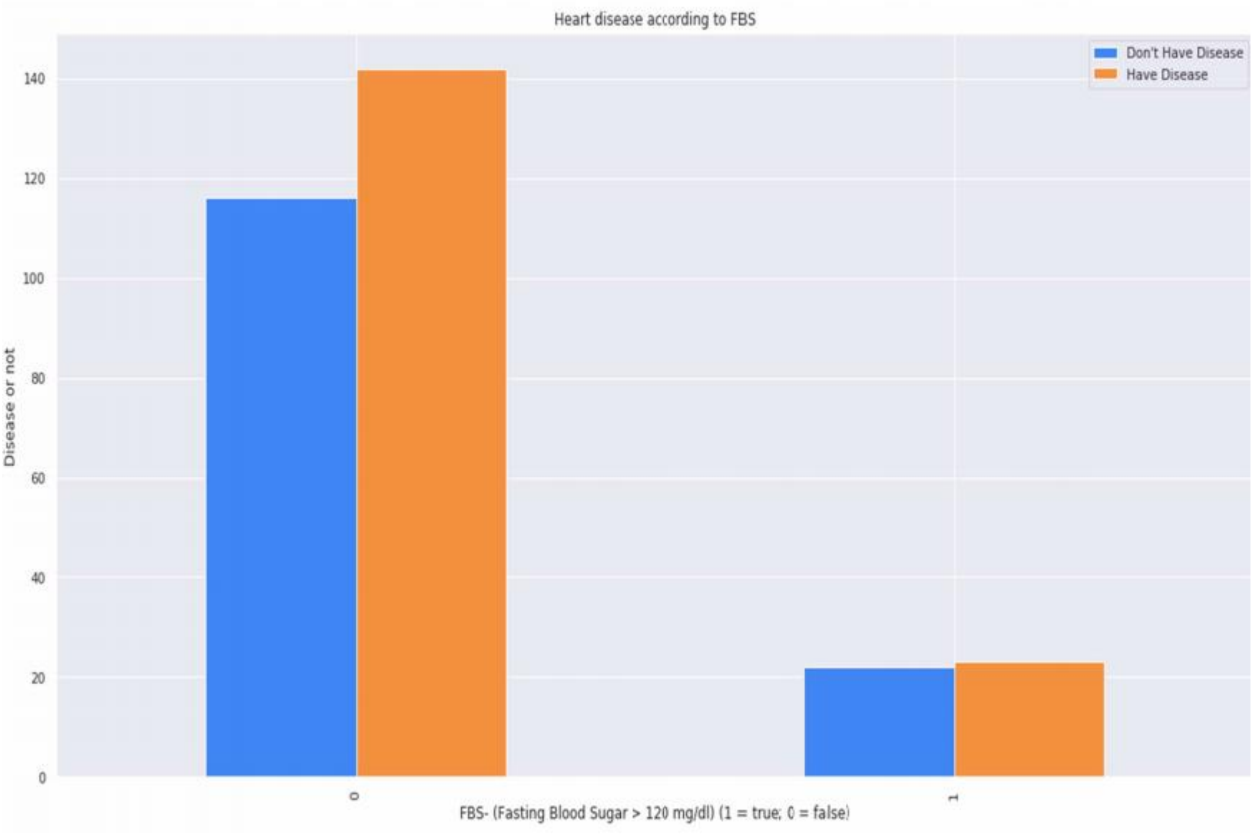
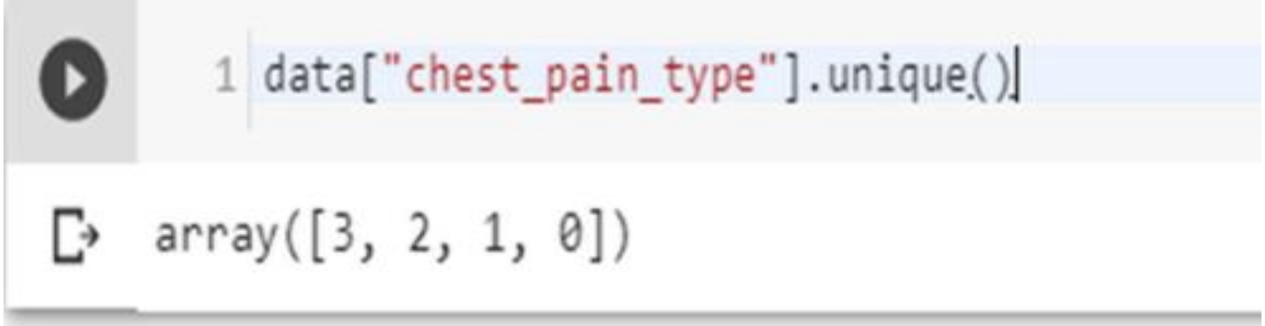


Figure 35 -Heart Disease according to FBS

3.11.8 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

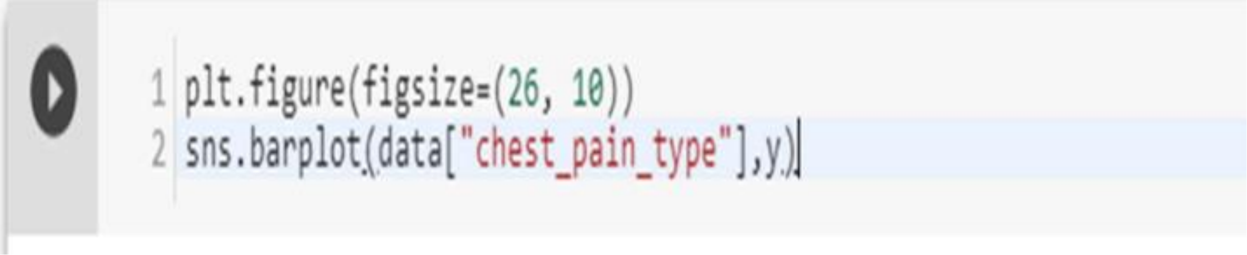


```
data["chest_pain_type"].unique()
```

```
array([3, 2, 1, 0])
```

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

Let’s plot chest pain types against target



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

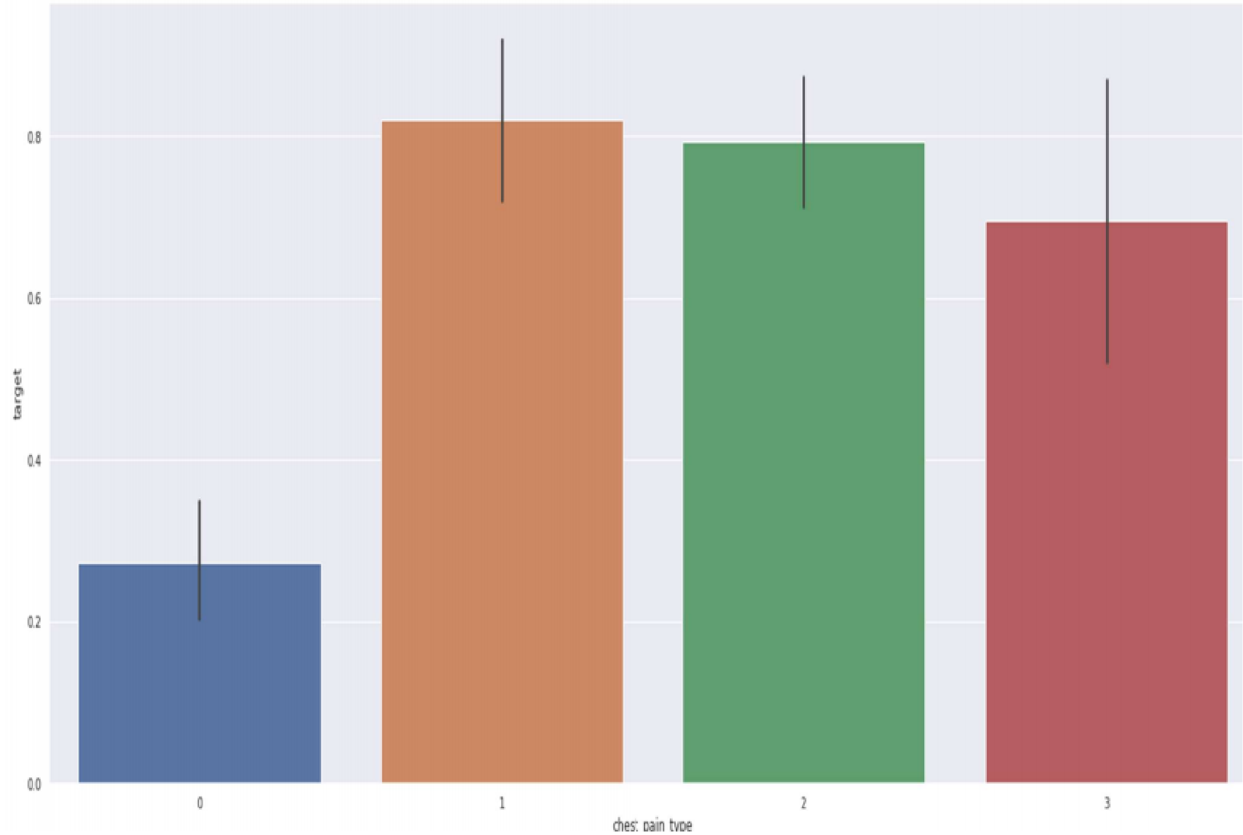
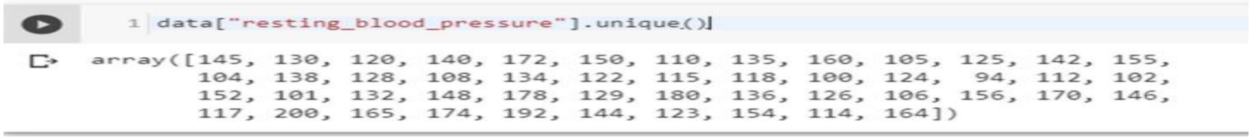


Figure 36 -Chest Pain Analysis

3.11.9 Analyzing the resting blood pressure

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

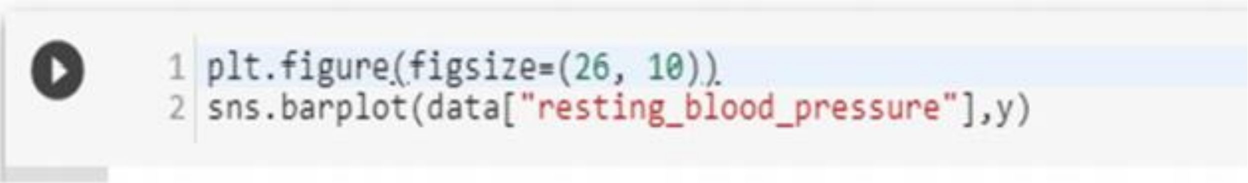
Let’s check out the unique resting blood pressures in our dataset



```
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



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

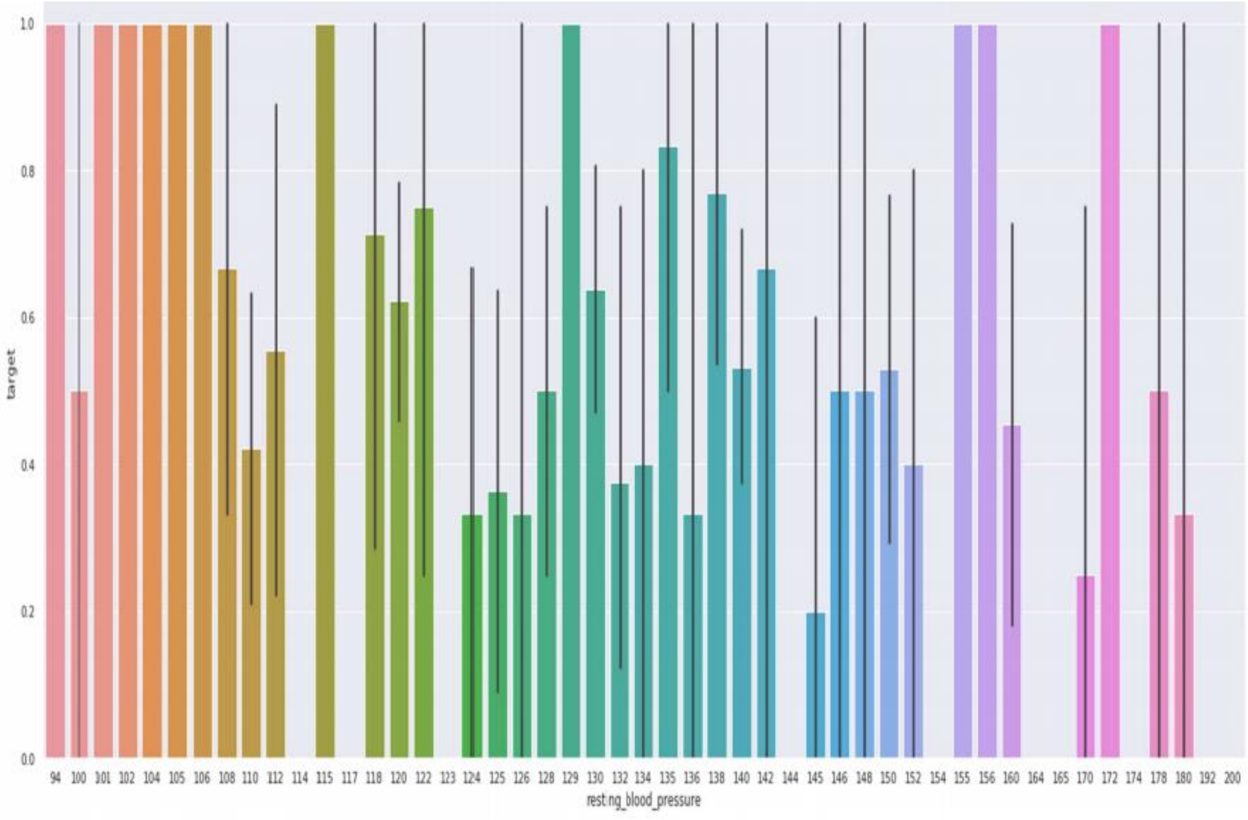
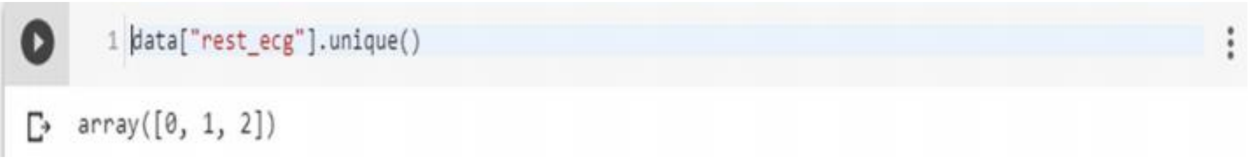


Figure 37 - Resting Blood Pressure Analysis

3.11.10 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

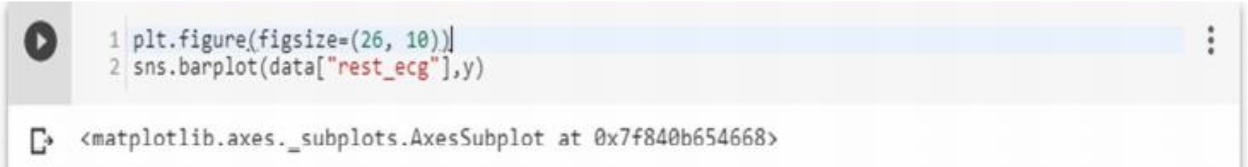


```
data["rest_ecg"].unique()
```

```
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



```
plt.figure(figsize=(26, 10))
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

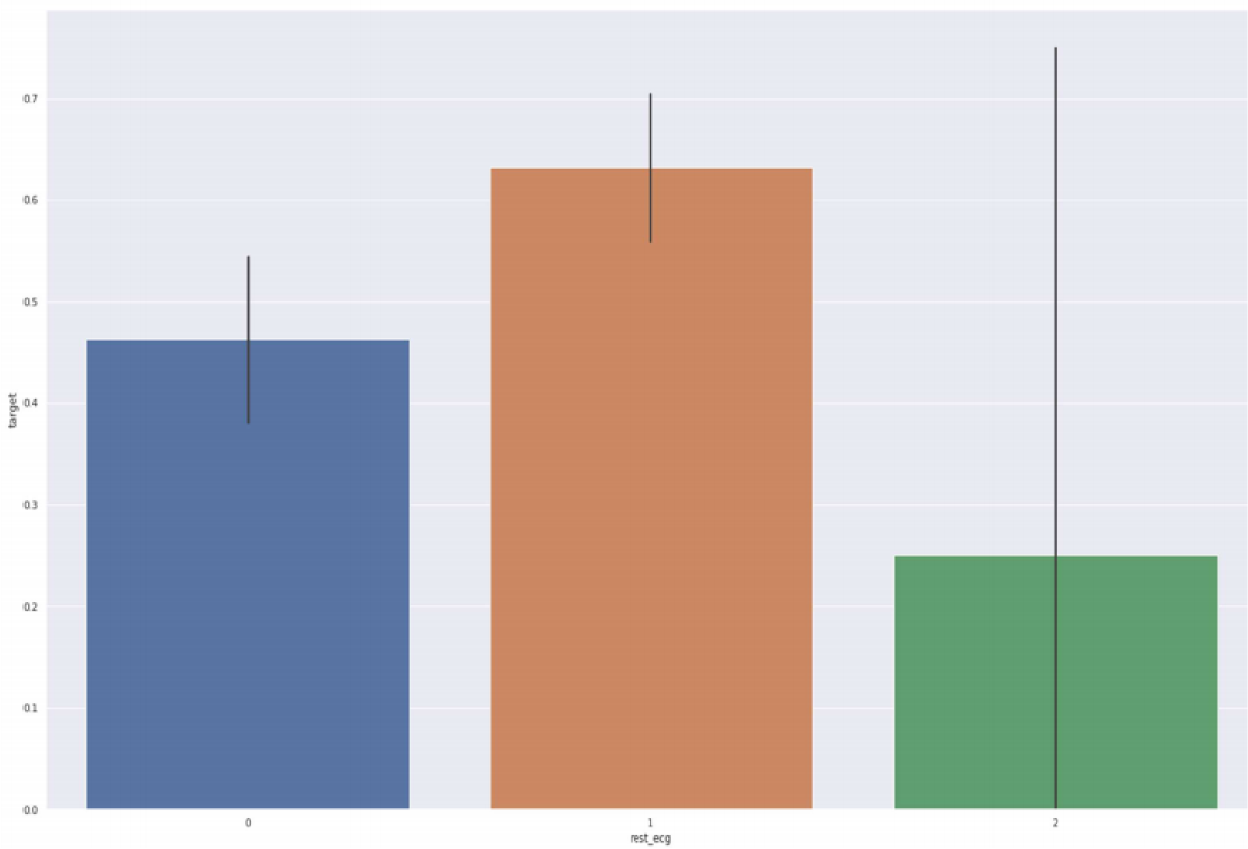
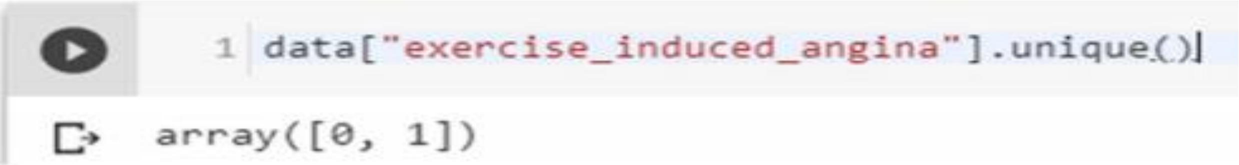


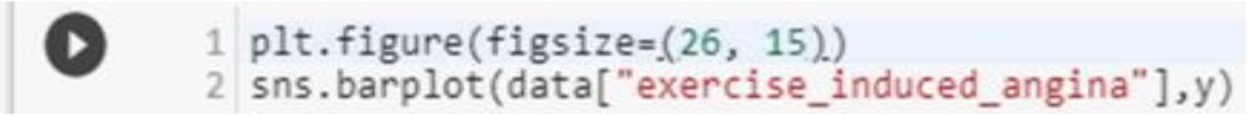
Figure 38 - ECG analysis

3.11.11 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)
```

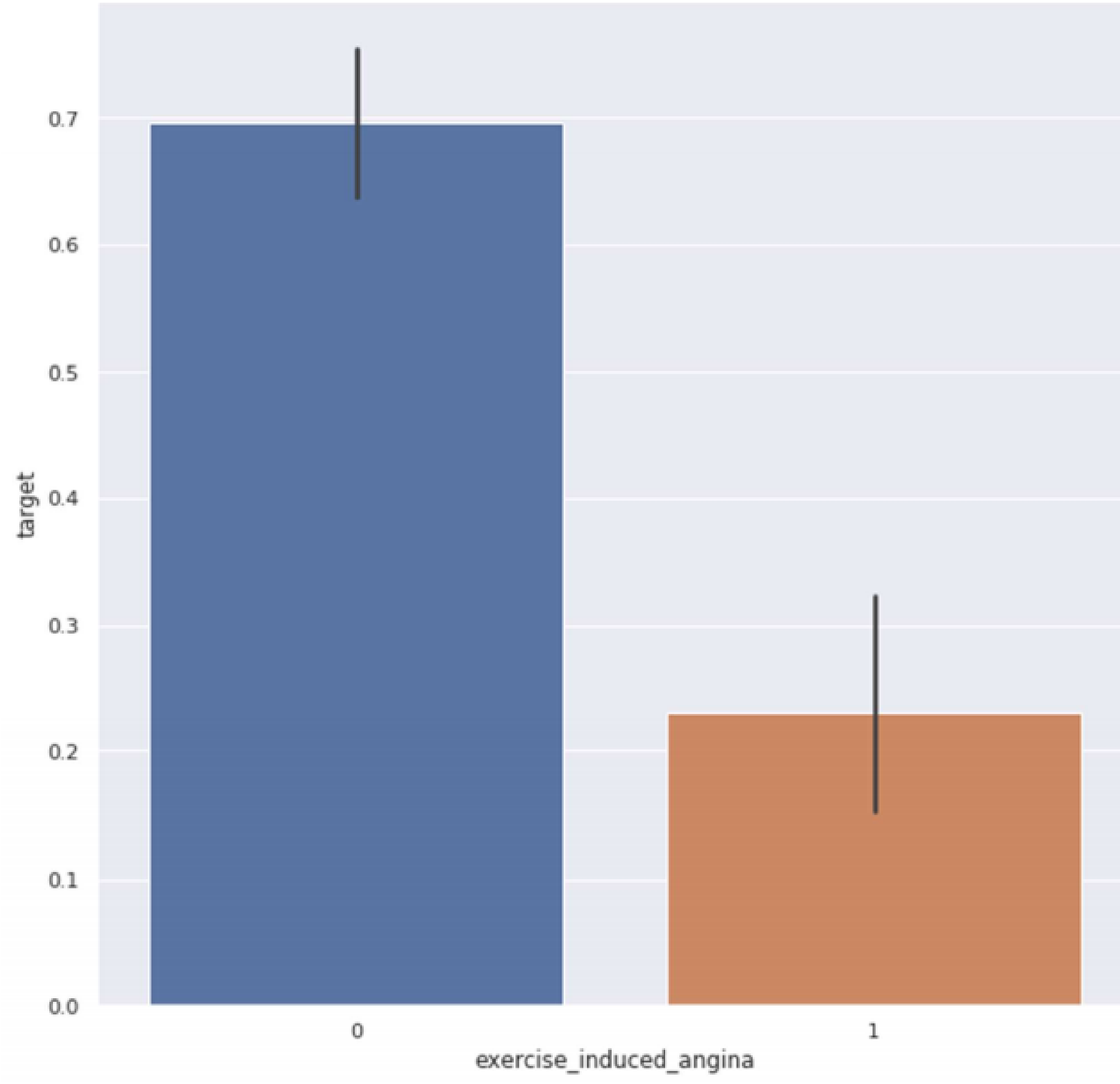
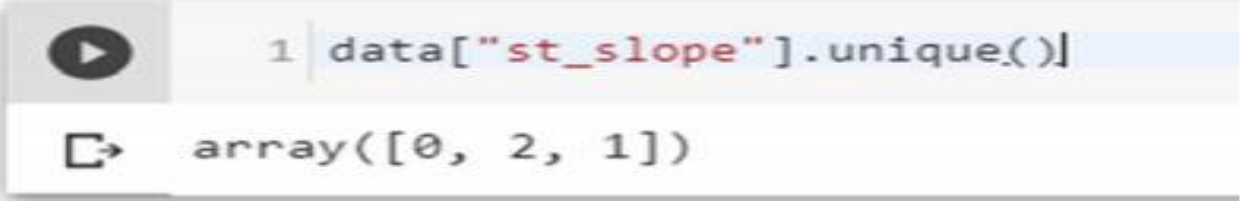


Figure 39 - Exercised Induced Angina Analysis

3.11.12 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.

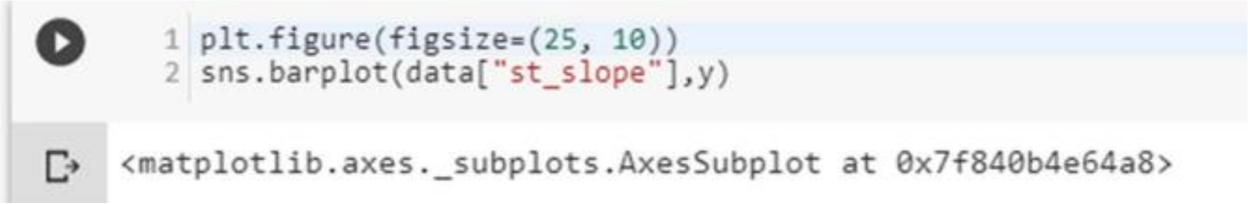
We herein present three cases, with evidence of ischemia on the treadmill ECG stress test. In addition, we discuss the use of the treadmill ECG stress test, including its indications, contraindications, reasons for termination and interpretation of the ST-segment changes, heart rate, as well as blood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



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

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

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



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

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

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

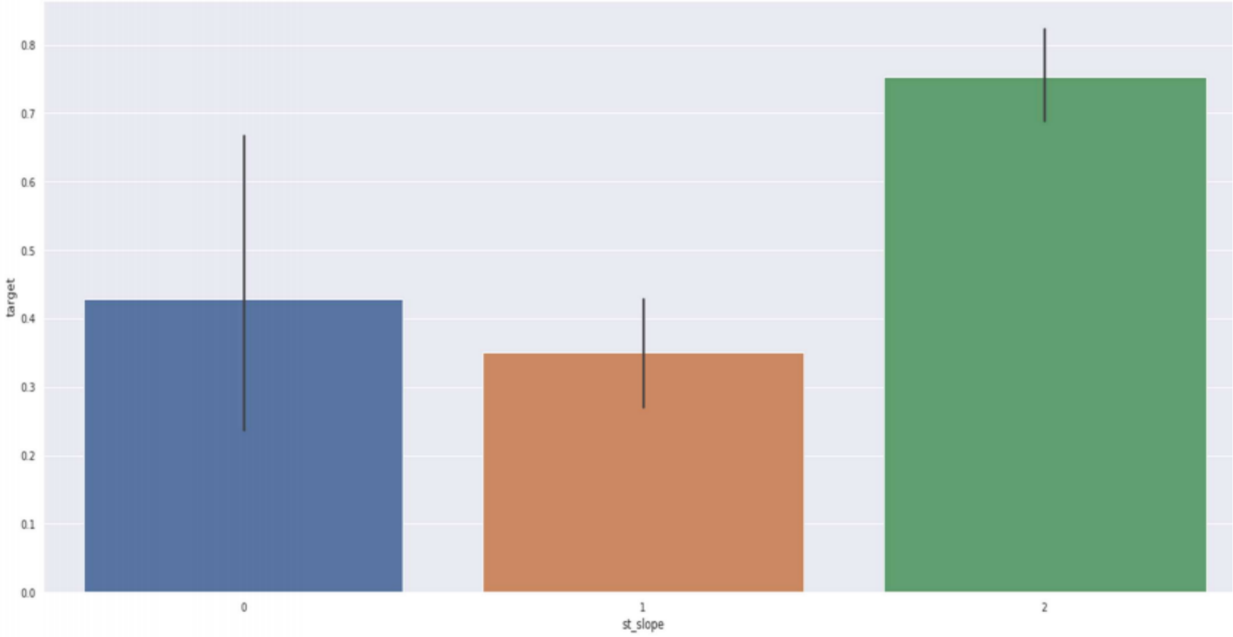


Figure 40- ST Slope Analysis

3.11.13 Analyzing no. of major vessels colored by fluoroscopy

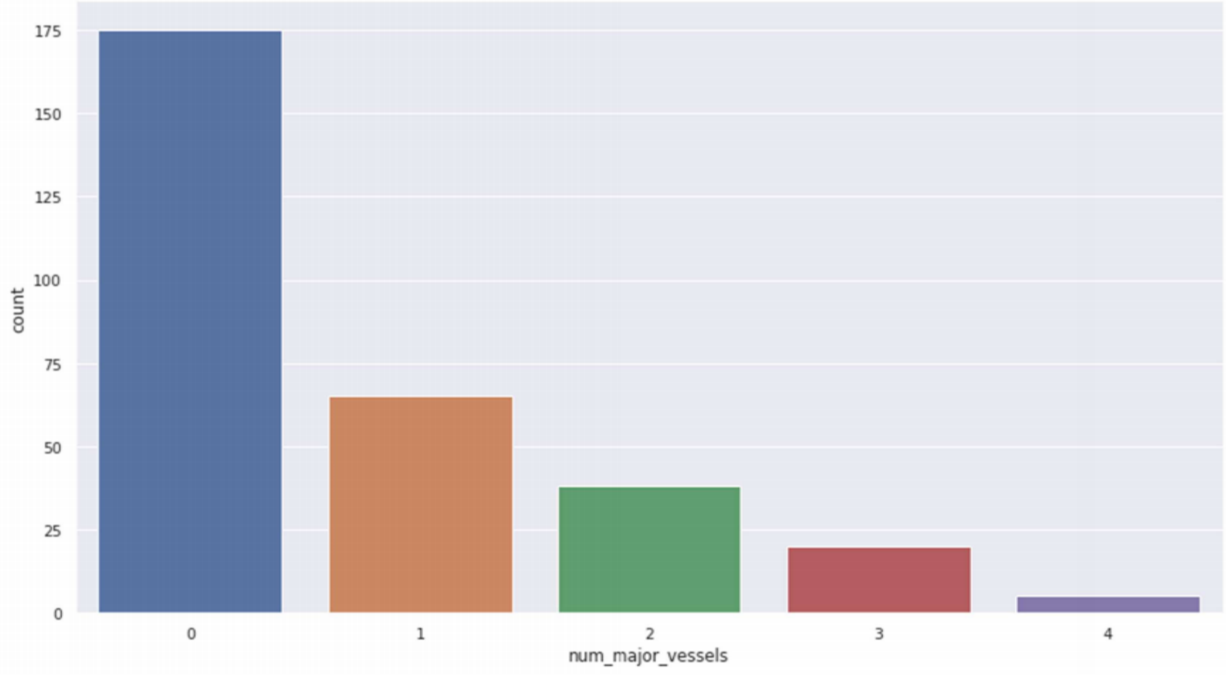


Figure 41- Analyzing no. of major vessels colored by fluoroscopy

3.11.13.1 Comparing with target:

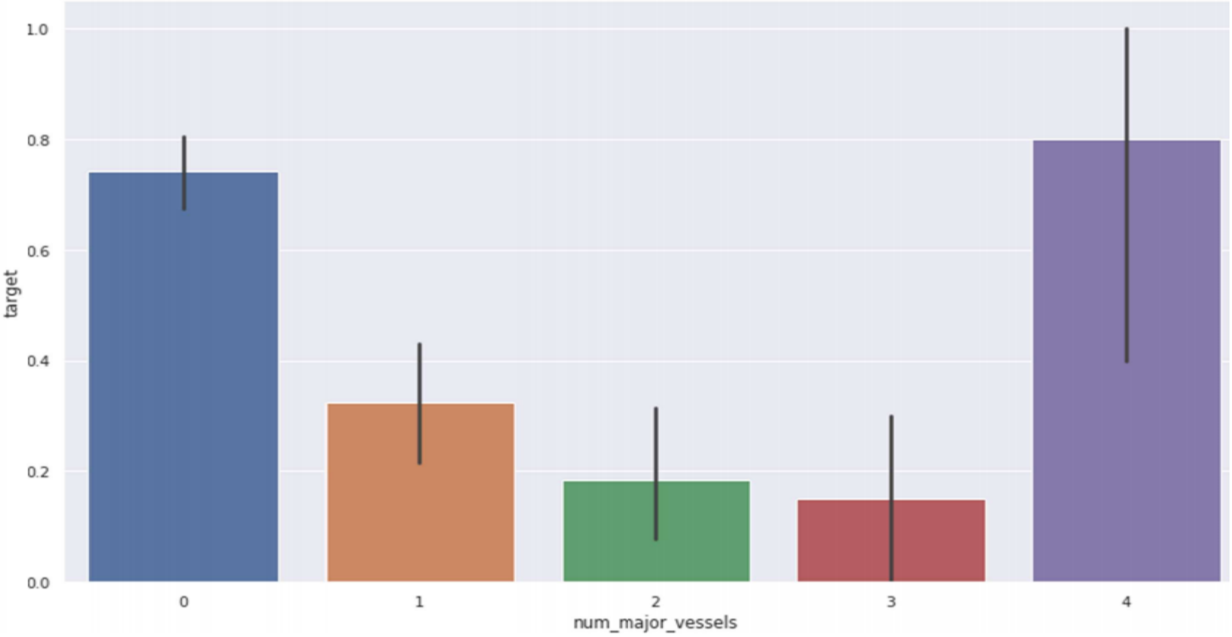


Figure 42 - Comparing with targets

3.11.14 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

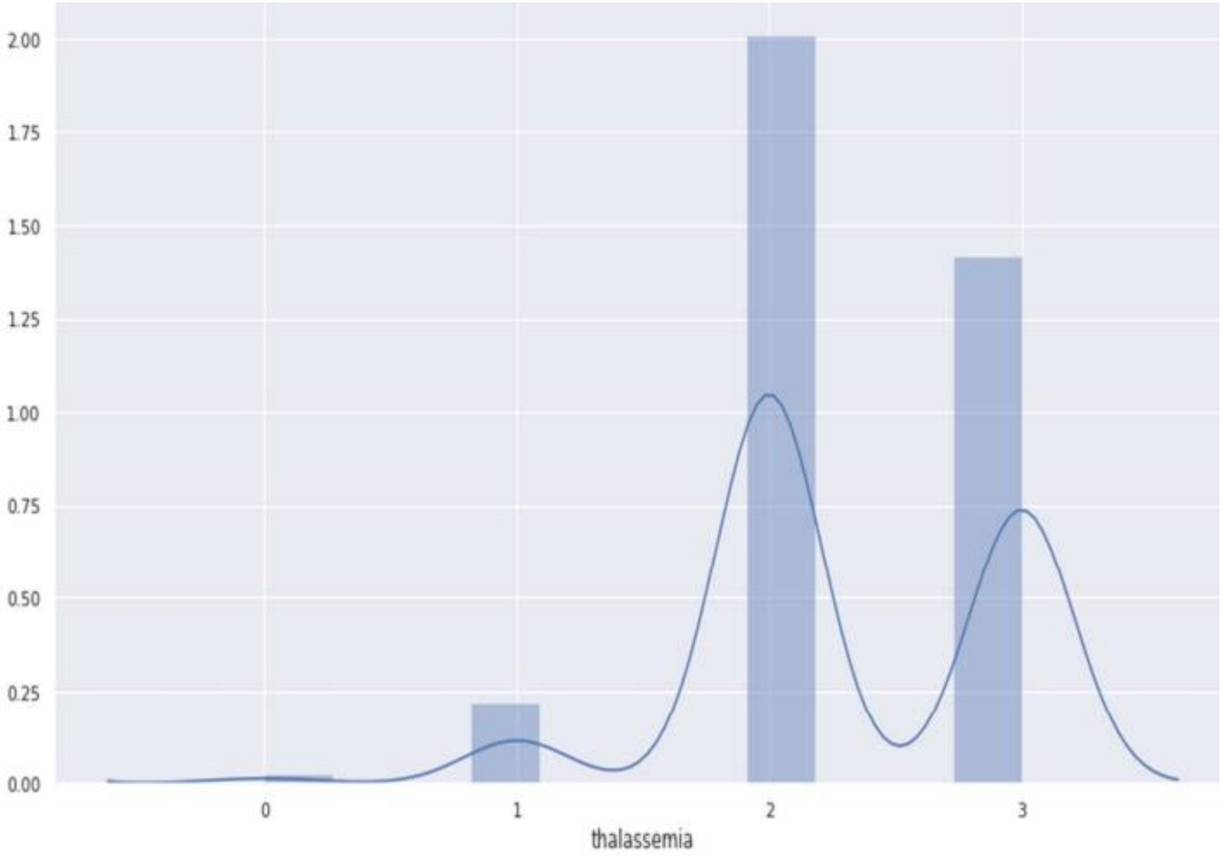


Figure 43-Thalassemia distribution

Against target:

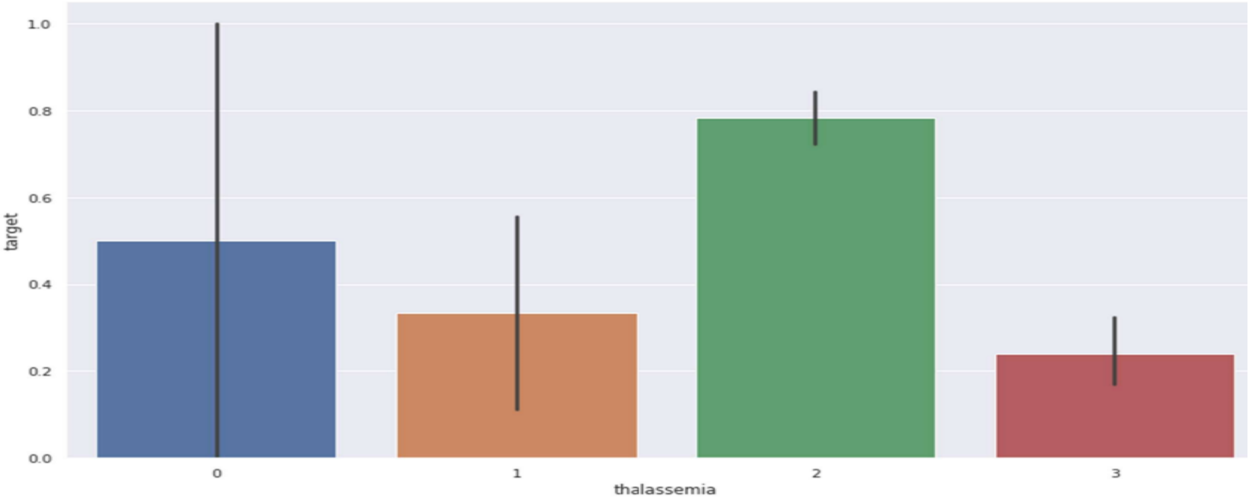


Figure 45 - Thalassemia against target

3.11.15 Thalassemia vs cholesterol

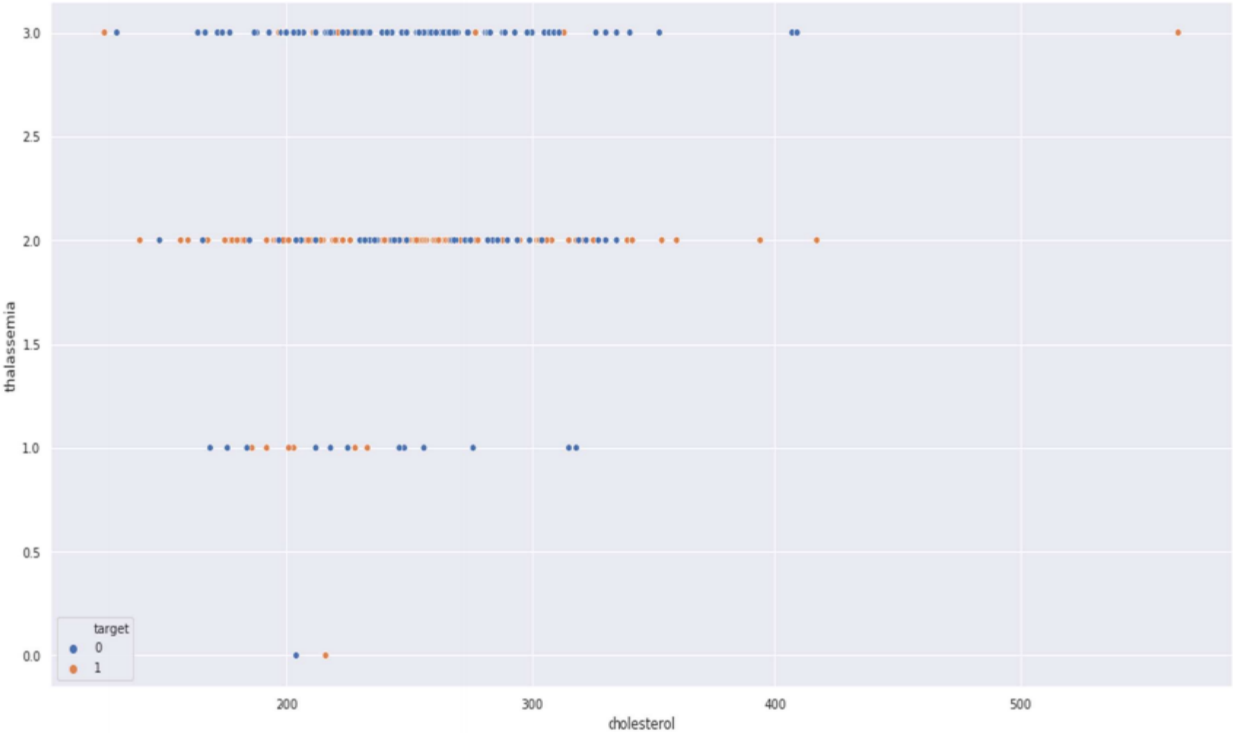


Figure 45 -Thalassemia vs cholesterol

Thalassemia vs resting blood pressure

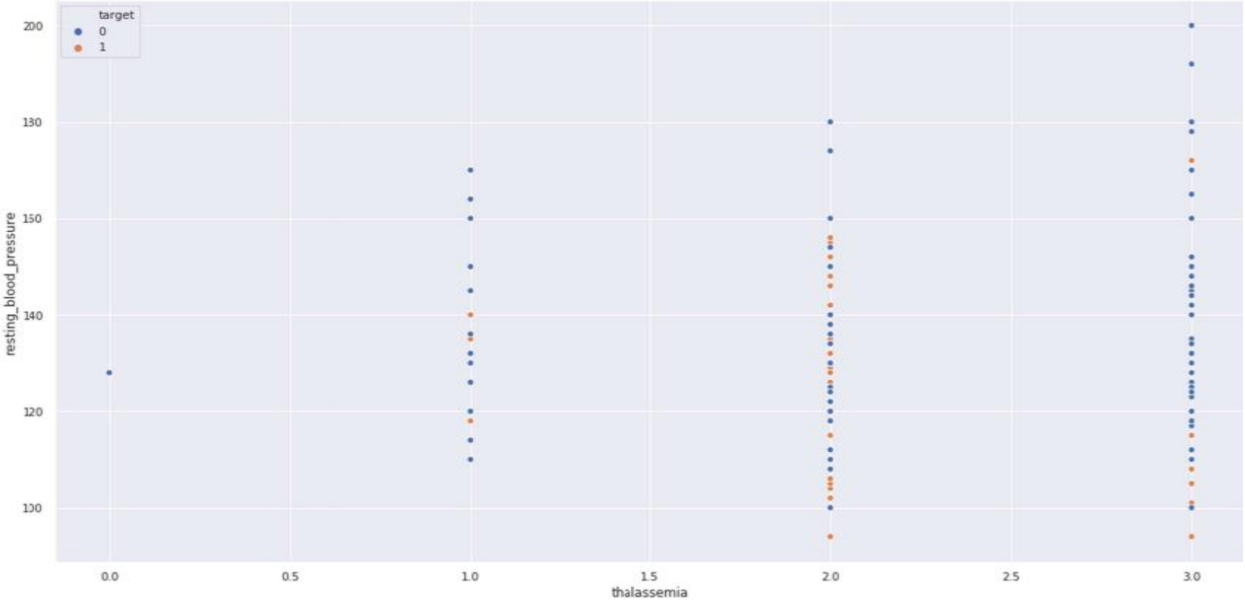
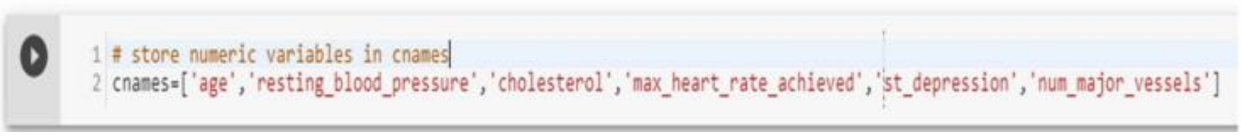


Figure 46 - Thalassemia vs resting blood pressure

3.12 Correlation Matrix

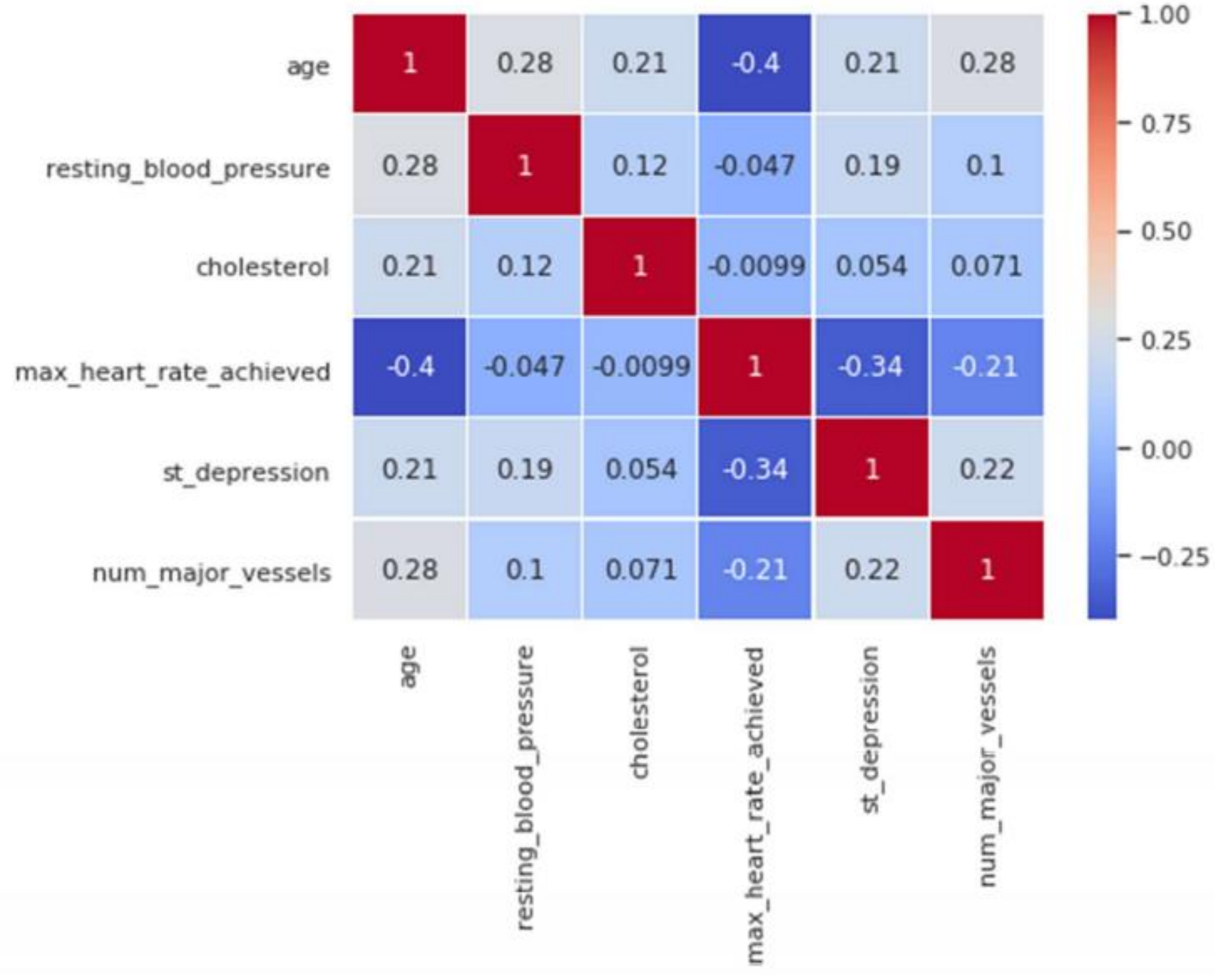
Correlation analysis is a method of statistical evaluation used to study the strength of a relationship between two, numerically measured, continuous variables (e.g. height and weight)

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']
```

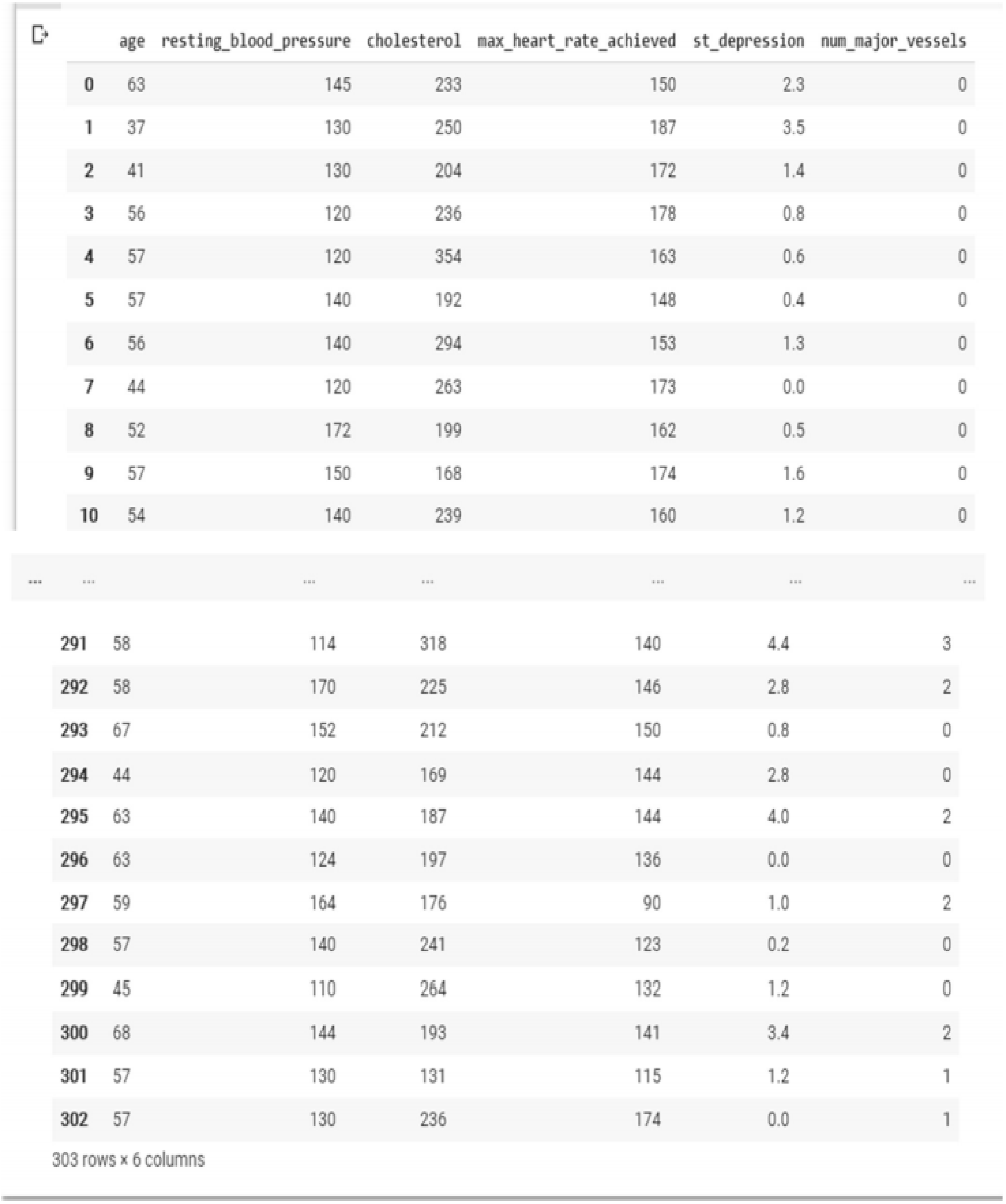
Let's plot the Correlation plot



	age	resting_blood_pressure	cholesterol	max_heart_rate_achieved	st_depression	num_major_vessels
age	1	0.28	0.21	-0.4	0.21	0.28
resting_blood_pressure	0.28	1	0.12	-0.047	0.19	0.1
cholesterol	0.21	0.12	1	-0.0099	0.054	0.071
max_heart_rate_achieved	-0.4	-0.047	-0.0099	1	-0.34	-0.21
st_depression	0.21	0.19	0.054	-0.34	1	0.22
num_major_vessels	0.28	0.1	0.071	-0.21	0.22	1

Figure 47 - Correlation Matrix

Correlation Analysis:



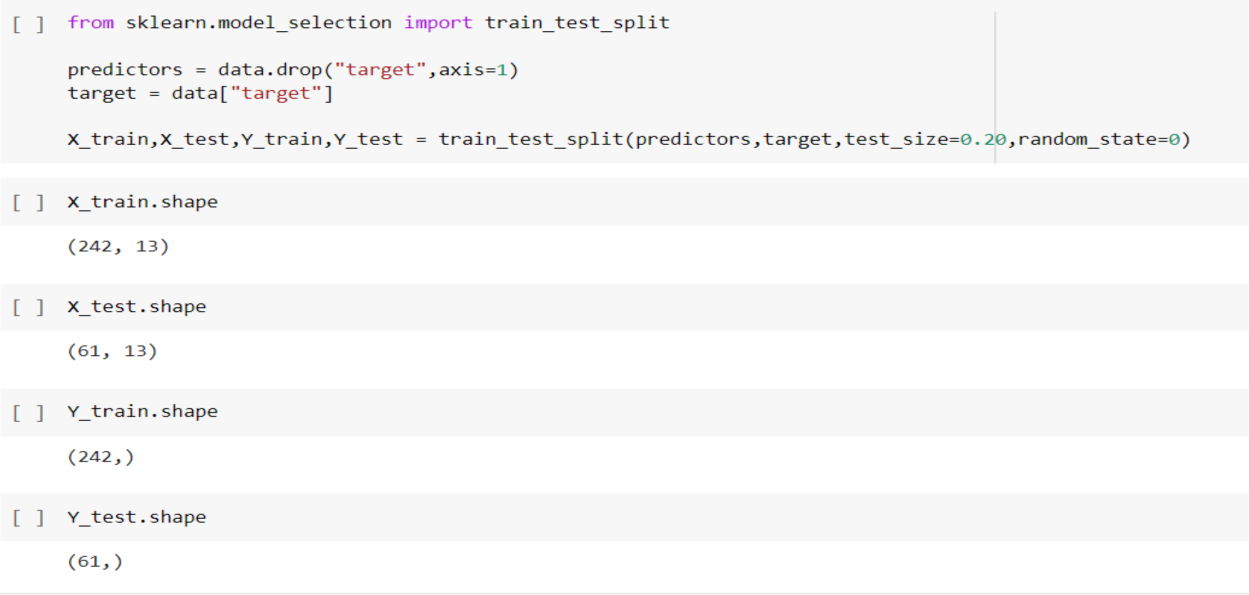
	age	resting_blood_pressure	cholesterol	max_heart_rate_achieved	st_depression	num_major_vessels
0	63	145	233	150	2.3	0
1	37	130	250	187	3.5	0
2	41	130	204	172	1.4	0
3	56	120	236	178	0.8	0
4	57	120	354	163	0.6	0
5	57	140	192	148	0.4	0
6	56	140	294	153	1.3	0
7	44	120	263	173	0.0	0
8	52	172	199	162	0.5	0
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.

Split the dataset to test and train

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

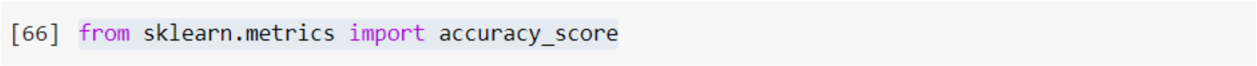


```
[ ]  from sklearn.model_selection import train_test_split

     predictors = data.drop("target",axis=1)
     target = data["target"]

     X_train,X_test,Y_train,Y_test = train_test_split(predictors,target,test_size=0.20,random_state=0)

[ ]  X_train.shape

     (242, 13)

[ ]  X_test.shape

     (61, 13)

[ ]  Y_train.shape

     (242,)

[ ]  Y_test.shape

     (61,)
```

3.8 Modeling 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).



```
[66]  from sklearn.metrics import accuracy_score
```

3.8.1 Naives 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.

$$\mathbf{MAP(h) = max(P(d|h) * P(h))}$$

3.8.2 Gaussian Naïve Bayes

$$\mathbf{mean(x) = 1/n * sum(x)}$$

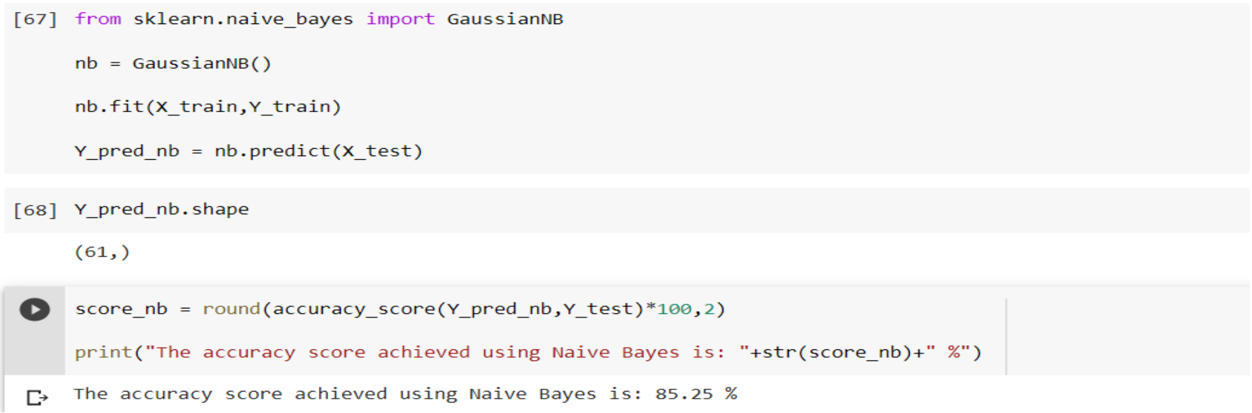
Where n is the number of instances and x are the values for an input variable in your training data. We can calculate the standard deviation using the following equation:

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

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.

$$\mathbf{pdf\ (x, mean, sd) = (1 / (sqrt\ (2 * PI) * sd)) * exp\ (\text{-}((x\text{-}mean\text{^}2)/(2*sd\text{^}2)))}$$

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



```
[67] from sklearn.naive_bayes import GaussianNB

     nb = GaussianNB()

     nb.fit(X_train,Y_train)

     Y_pred_nb = nb.predict(X_test)
```

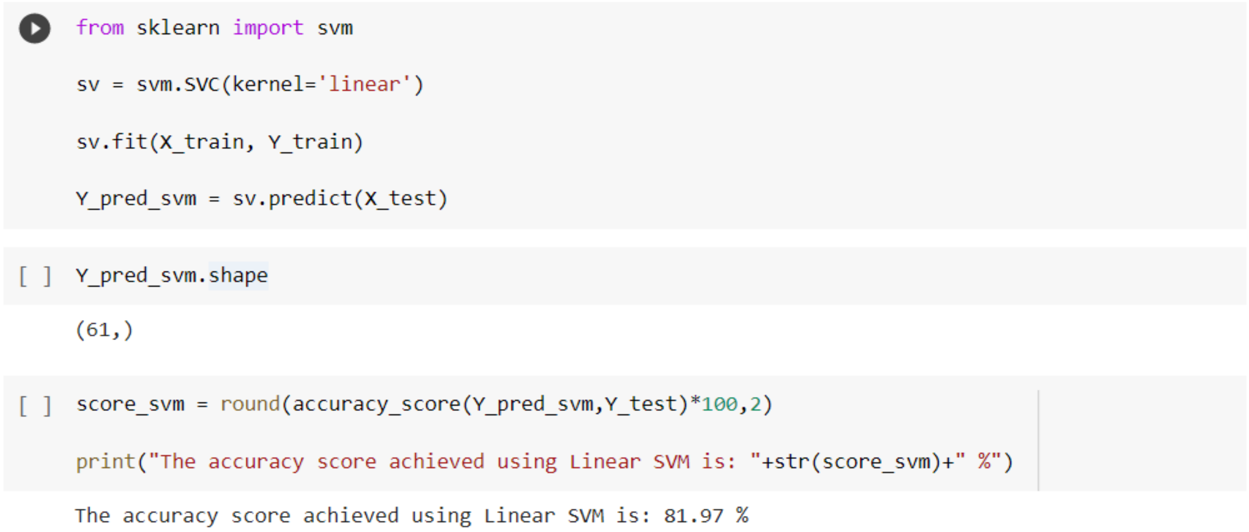
```
[68] Y_pred_nb.shape

     (61,)
```

```
score_nb = round(accuracy_score(Y_pred_nb,Y_test)*100,2)

print("The accuracy score achieved using Naive Bayes is: "+str(score_nb)+" %")

The accuracy score achieved using Naive Bayes is: 85.25 %
```

3.8.3 Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.



```
from sklearn import svm

sv = svm.SVC(kernel='linear')

sv.fit(X_train, Y_train)

Y_pred_svm = sv.predict(X_test)
```

```
[ ] Y_pred_svm.shape

(61,)
```

```
[ ] score_svm = round(accuracy_score(Y_pred_svm,Y_test)*100,2)

    print("The accuracy score achieved using Linear SVM is: "+str(score_svm)+" %")

The accuracy score achieved using Linear SVM is: 81.97 %
```

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

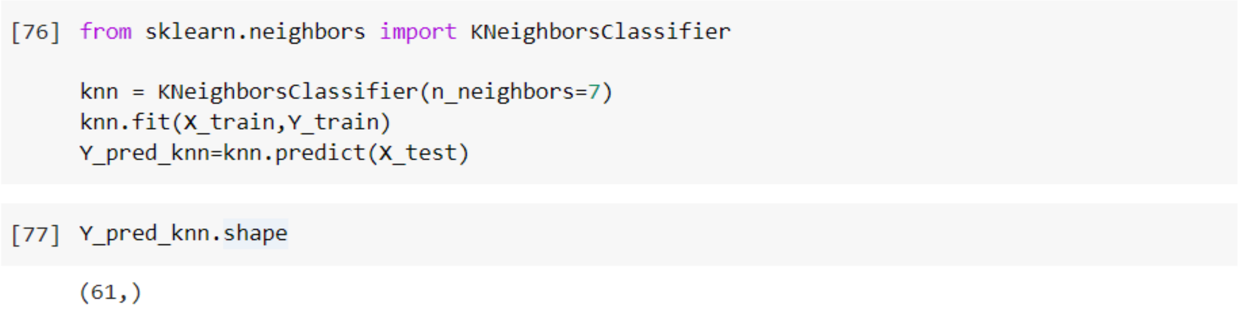
• Calculate the distance between test data and each row of training data. Here we will use Euclidean distance as our distance metric since it’s the most popular method. The other metrics that can be used are Chebyshev, cosine, etc.

• Sort the calculated distances in ascending order based on distance values

- Get top k rows from the sorted array

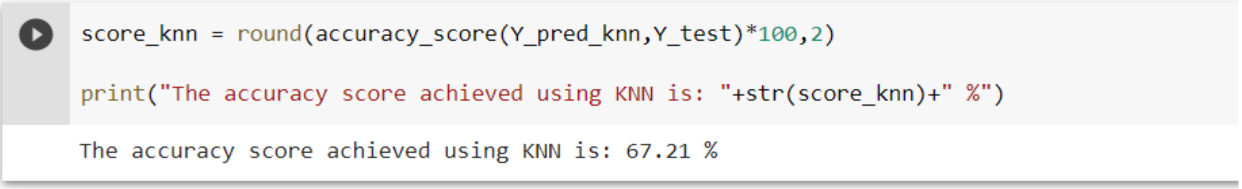
- Get the most frequent class of these rows

- Return the predicted class



```
[76] from sklearn.neighbors import KNeighborsClassifier

     knn = KNeighborsClassifier(n_neighbors=7)
     knn.fit(X_train,Y_train)
     Y_pred_knn=knn.predict(X_test)
```

```
[77] Y_pred_knn.shape

     (61,)
```



```
score_knn = round(accuracy_score(Y_pred_knn,Y_test)*100,2)

print("The accuracy score achieved using KNN is: "+str(score_knn)+" %")
```

```
The accuracy score achieved using KNN is: 67.21 %
```

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.

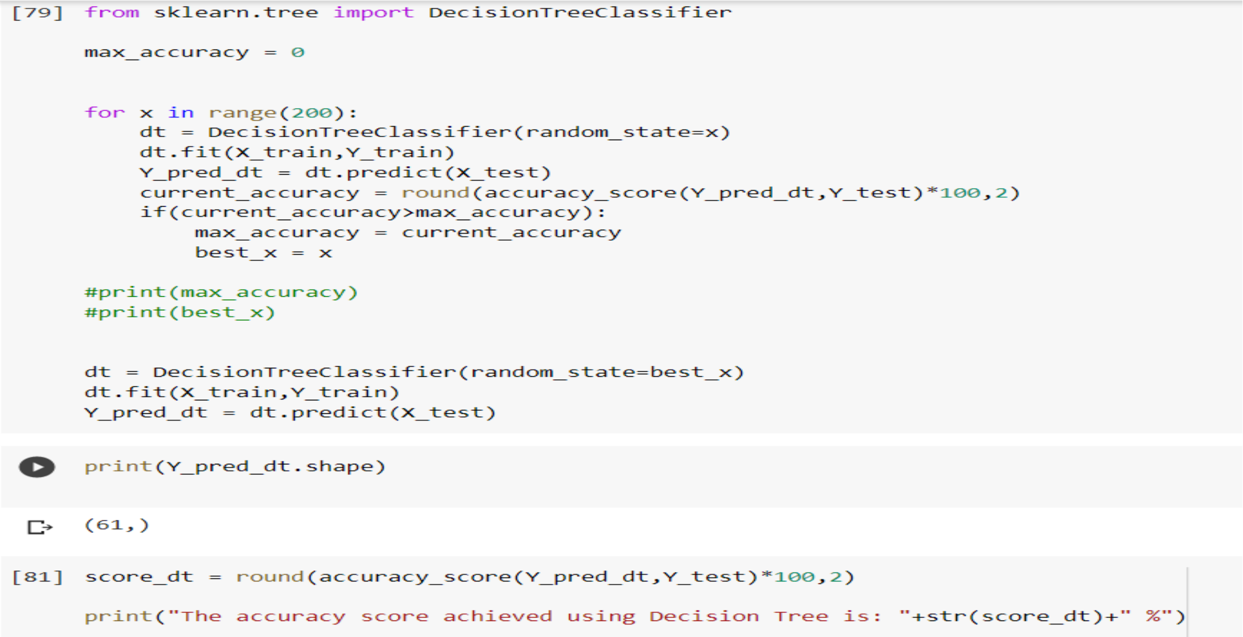
- Order to placing attributes as root or internal node of the tree is done by using some statistical approach. The popular attribute selection measures

- 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.



```
[79] from sklearn.tree import DecisionTreeClassifier

     max_accuracy = 0


     for x in range(200):
         dt = DecisionTreeClassifier(random_state=x)
         dt.fit(X_train,Y_train)
         Y_pred_dt = dt.predict(X_test)
         current_accuracy = round(accuracy_score(Y_pred_dt,Y_test)*100,2)
         if(current_accuracy>max_accuracy):
             max_accuracy = current_accuracy
             best_x = x

     #print(max_accuracy)
     #print(best_x)


     dt = DecisionTreeClassifier(random_state=best_x)
     dt.fit(X_train,Y_train)
     Y_pred_dt = dt.predict(X_test)
```

```
print(Y_pred_dt.shape)
```

```
(61,)
```

```
[81] score_dt = round(accuracy_score(Y_pred_dt,Y_test)*100,2)

     print("The accuracy score achieved using Decision Tree is: "+str(score_dt)+" %")
```

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.

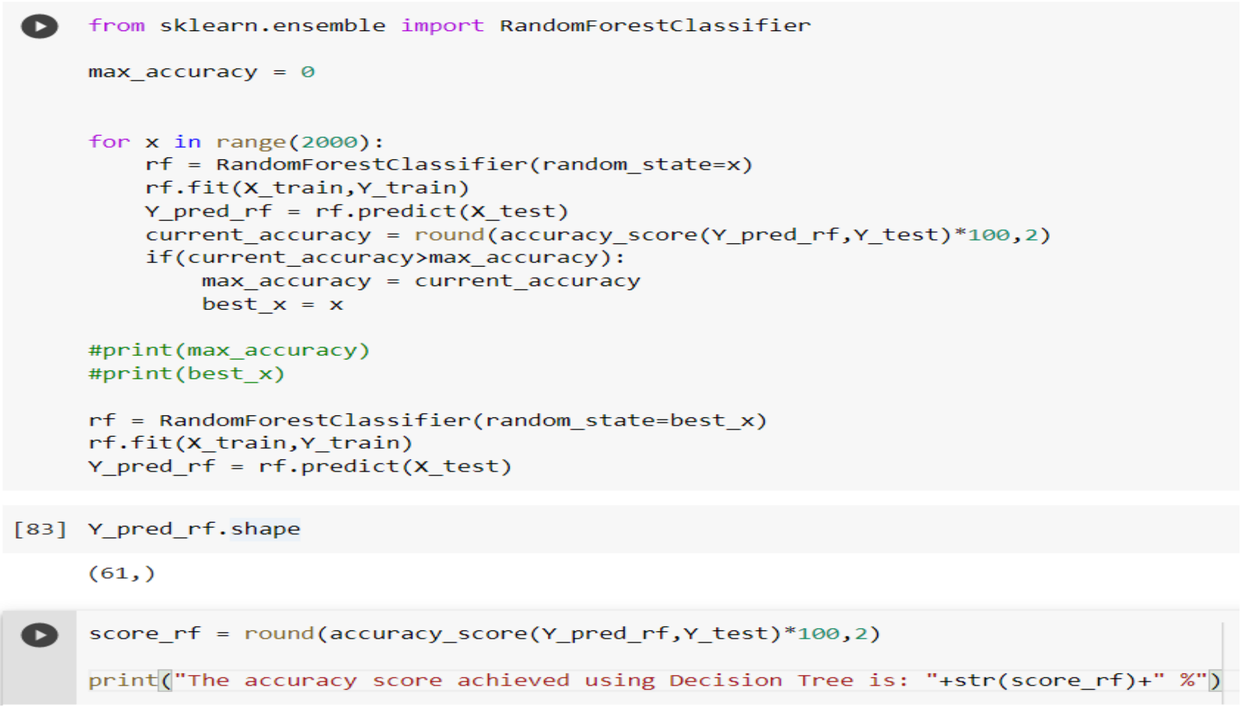
5. Build forest by repeating steps 1 to 4 for “n” number times to create “n” number of trees.

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.

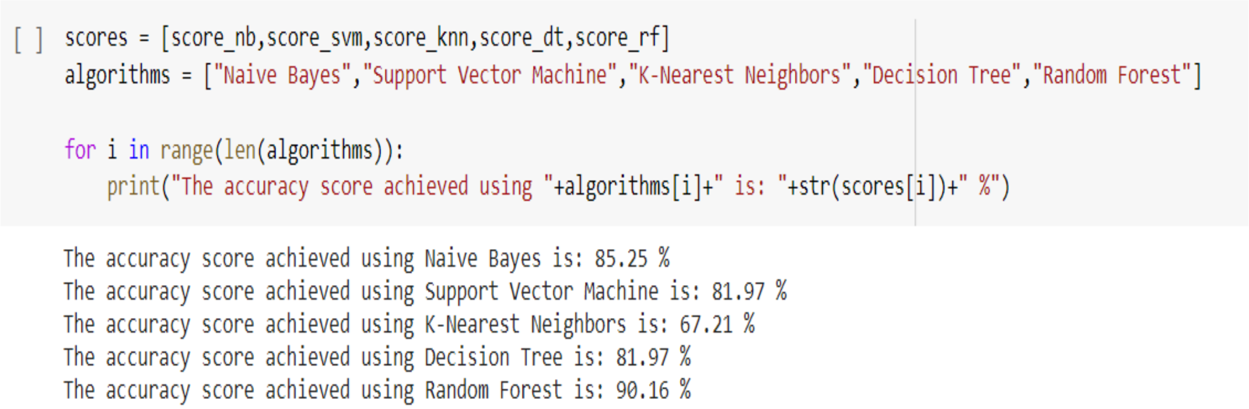


```
from sklearn.ensemble import RandomForestClassifier

max_accuracy = 0


for x in range(2000):
    rf = RandomForestClassifier(random_state=x)
    rf.fit(X_train,Y_train)
    Y_pred_rf = rf.predict(X_test)
    current_accuracy = round(accuracy_score(Y_pred_rf,Y_test)*100,2)
    if(current_accuracy>max_accuracy):
        max_accuracy = current_accuracy
        best_x = x

#print(max_accuracy)
#print(best_x)

rf = RandomForestClassifier(random_state=best_x)
rf.fit(X_train,Y_train)
Y_pred_rf = rf.predict(X_test)
```

```
[83] Y_pred_rf.shape

     (61,)
```

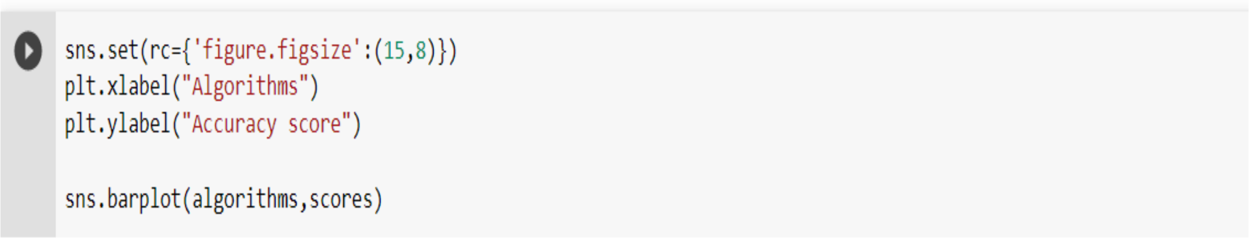
```
score_rf = round(accuracy_score(Y_pred_rf,Y_test)*100,2)

print("The accuracy score achieved using Decision Tree is: "+str(score_rf)+" %")
```

Final output



```
[ ]  scores = [score_nb,score_svm,score_knn,score_dt,score_rf]
     algorithms = ["Naive Bayes","Support Vector Machine","K-Nearest Neighbors","Decision Tree","Random Forest"]

     for i in range(len(algorithms)):
         print("The accuracy score achieved using "+algorithms[i]+" is: "+str(scores[i])+" %")
```

```
The accuracy score achieved using Naive Bayes is: 85.25 %
The accuracy score achieved using Support Vector Machine is: 81.97 %
The accuracy score achieved using K-Nearest Neighbors is: 67.21 %
The accuracy score achieved using Decision Tree is: 81.97 %
The accuracy score achieved using Random Forest is: 90.16 %
```



```
sns.set(rc={'figure.figsize':(15,8)})
plt.xlabel("Algorithms")
plt.ylabel("Accuracy score")

sns.barplot(algorithms,scores)
```

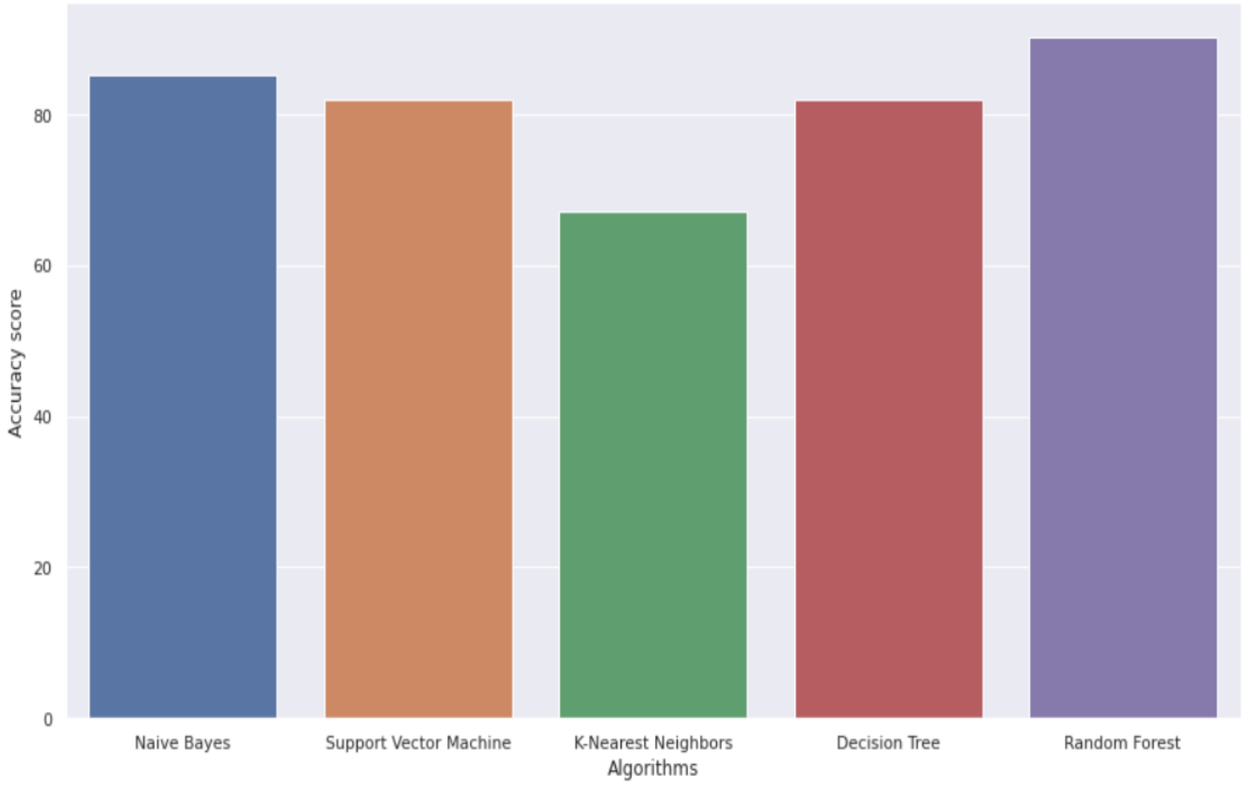


Figure 48 Accuracy Scores of all Algorithms

5.3 ALGORITHMS USED

- Naïve Bayes
- Support Vector Machine
- K – Nearest Neighbour
- Decision Tree
- Random Forest

8.1.Naïve Bayes

Naive Bayes is a simple but an effective classification technique which is based on the Bayes Theorem. It assumes independence among predictors, i.e., the attributes or features should be not correlated to one another or should not, in anyway, be related to each other. Even if there is dependency, still all these features or attributes independently contribute to the probability and that is why it is called Naïve. In [7], Naive Bayes has achieved an accuracy of 84.1584% with the 10 most significant features which are selected using SVM-RFE (Recursive Feature Elimination) and gain ratio algorithms whereas in[8],Naive Bayes has achieved an accuracy of 83.49% when all 13 attributes of the Cleveland dataset[25] are used.

8.2.Support Vector Machine

Support Vector Machine is an extremely popular supervised machine learning technique(having a pre-defined target variable) which can be used as a classifier as well as a predictor. For classification, it finds a hyper-plane in the feature space that differentiates between the classes. An SVM model represents the training data points as points in the feature space, mapped in such a way that points belonging to separate classes are segregated by a margin as wide as possible. The test data points are then mapped into that same space and are classified based on which side of the margin they fall.

8.3. K – Nearest Neighbour

In 1951, Hodges et al. introduced a nonparametric technique for pattern classification which is popularly known the K-Nearest Neighbour rule[13]. K-Nearest Neighbour technique is one of the most elementary but very effective classification techniques. It makes no assumptions about the data and is generally be used for classification tasks when there is very less or no prior knowledge about the data distribution. This algorithm involves finding the k nearest data points in the training set to the data point for which a target value is unavailable and assigning the average value of the found data points to it. In KNN gives an accuracy of 83.16% when the value of k is equal to 9 while using 10-cross validation technique. In KNN with Ant Colony Optimization performs better than other techniques with an accuracy of 70.26% and the error rates is 0.526.Ridhi Saini et al. have obtained a efficiency of 87.5% [15], which is very good.

8.4. Decision tree

Decision tree is a of supervised learning algorithm.This technique is mostly used in classification problems. It performs effortlessly with continuous and categorical attributes. This algorithm divides the population into two or more similar sets based on the most Significant predictors. Decision Tree algorithm, first calculates the entropy of each and every attribute. Then the dataset is split with the help of the variables or predictors with maximum information gain or minimum entropy. These two steps are performed recursively with the remaining attributes. In [10]decision tree has the worst performance with an accuracy of 77.55% but when decision tree is used with boosting technique it performs better with an accuracy of 82.17%.In [9] decision tree performs very poorly with a correctly classified instance percentage of 42.8954% whereas in [16] also uses the same dataset but used the J48 algorithm for implementing Decision Trees and the accuracy thus obtained is 67.7% which is less but still an improvement on the former.

8.5.Random Forest

Random Forest is also a popularly supervised machine learning algorithm. This technique can be used for both regression and classification tasks but generally performs better in classification tasks. As the name suggests, Random Forest technique considers multiple decision trees before giving an output. So, it is basically an ensemble of decision trees. This technique is based on the belief that more number of trees would converge to the right decision. For classification, it uses a voting system and then decides the class whereas in regression it takes the mean of all the outputs of each of the decision trees. It works well with large datasets with high dimensionality

5.4 SAMPLE CODE

```
# I. Importing essential 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')
# II. Importing and understanding our dataset
data = pd.read_csv("heart.csv")
# Verifying it as a 'dataframe' object in pandas
type(data)
# Shape of dataset
data.shape
# Printing out a few columns
data.head(5)
data.sample(5)
# Description
data.describe()
```

```
# Data Preprocessing

data.info()

data.isnull().sum()

# So, we have no missing values

# Let's understand our columns better:

info = ["age","1: male, 0: female","chest pain type, 1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic","resting blood pressure"," serum cholestoral in mg/dl","fasting blood sugar > 120 mg/dl","resting electrocardiographic results (values 0,1,2)"," maximum heart rate achieved","exercise induced angina","oldpeak = ST depression induced by exercise relative to rest","the slope of the peak exercise ST segment","number of major vessels (0-3) colored by flourosopy","thal: 3 = normal; 6 = fixed defect; 7 = reversable defect"]

for i in range(len(info)):

    print(data.columns[i]+":\t\t\t"+info[i])

# Analysing the 'target' variable

data["target"].describe()

data["target"].unique()

# Checking correlation between columns

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

# III Exploratory Data Analysis (EDA)

# First, analysing the target variable

y = data["target"]

sns.countplot(y)

target_temp = data.target.value_counts()

print(target_temp)
```

```
# 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)))
# Analysing the 'Sex' feature
data["sex"].unique()
sns.barplot(data["sex"],y)
# count male and female patients
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))
# 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()
# 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()

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']

# Heart disease according to Fasting Blood sugar

pd.crosstab(data.fasting_blood_sugar,
data.target).plot(kind="bar",figsize=(20,10),color=['pink','#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()

# Analysing the chest pain (4 types of chest pain) [Value 1: typical angina, Value 2: atypical
angina, Value 3: non-anginal pain, Value 4: asymptomatic]

data["chest_pain_type"].unique()

plt.figure(figsize=(26, 10))

sns.barplot(data["chest_pain_type"],y)

# Analysing The person's resting blood pressure

data["resting_blood_pressure"].unique()

plt.figure(figsize=(26, 10))
```

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

# Analysing the Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)

data["rest_ecg"].unique()

plt.figure(figsize=(26, 15))

sns.barplot(data["rest_ecg"],y)

# Analysing Exercise induced angina

data["exercise_induced_angina"].unique()

plt.figure(figsize=(10, 10))

sns.barplot(data["exercise_induced_angina"],y)

# Analysing the slope of the peak exercise ST segment (Value 1: upsloping, Value 2: flat, Value 3: downsloping)

data["st_slope"].unique()

plt.figure(figsize=(25, 10))

sns.barplot(data["st_slope"],y)

# Analysing number of major vessels (0-3) colored by flourosopy

data["num_major_vessels"].unique()

# count num_major vessels

sns.countplot(data["num_major_vessels"])

# comparing with target

sns.barplot(data["num_major_vessels"],y)

# Analysing A blood disorder called thalassemia (1 = normal; 2 = fixed defect; 3  = reversable defect)
```

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

# plotting the thalassemia distribution (0,1,2,3)

sns.distplot(data["thalassemia"])

# comparing with target

sns.barplot(data["thalassemia"],y)

# thalassemia and cholesterol scatterplot

plt.figure(figsize=(20,10))

sns.scatterplot(x='cholesterol',y='thalassemia',data=data,hue='target')

plt.show()

# thalassemia vs resting blood pressure scatterplot

plt.figure(figsize=(20,10))

sns.scatterplot(x='thalassemia',y='resting_blood_pressure',data=data,hue='target')

plt.show()

# Health rate vs age

plt.figure(figsize=(20, 10))

plt.scatter(x=data.age[data.target==1], y=data.thalassemia[(data.target==1)], c="green")

plt.scatter(x=data.age[data.target==0], y=data.thalassemia[(data.target==0)])

plt.legend(["Disease", "Not Disease"])

plt.xlabel("Age")

plt.ylabel("Maximum Heart Rate")

plt.show()
```

```
# Correlation plot

# store numeric variables in cnames

cnames=['age','resting_blood_pressure','cholesterol','max_heart_rate_achieved','st_depression','num_major_vessels']

#Set the width and height of the plot

f, ax = plt.subplots(figsize=(7, 5))

# Correlation plot

df_corr = data.loc[:,cnames]

# Generate correlation matrix

corr = df_corr.corr()

#Plot using seaborn library

sns.heatmap(corr, annot = True, cmap='coolwarm',linewidths=.1)

plt.show()

# Correlation analysis

df_corr = data.loc[:,cnames]

df_corr

# IV Splitting the dataset to Train and Test

from sklearn.model_selection import train_test_split

predictors = data.drop("target",axis=1)

target = data["target"]

X_train,X_test,Y_train,Y_test = train_test_split(predictors,target,test_size=0.20,random_state=0)

X_train.shape

X_test.shape
```

```
Y_train.shape

Y_test.shape

# V Model Fitting

from sklearn.metrics import accuracy_score


# Naive Bayes

from sklearn.naive_bayes import GaussianNB

nb = GaussianNB()

nb.fit(X_train,Y_train)

Y_pred_nb = nb.predict(X_test)

Y_pred_nb.shape

score_nb = round(accuracy_score(Y_pred_nb,Y_test)*100,2)

print("The accuracy score achieved using Naive Bayes is: "+str(score_nb)+" %")


# SVM

from sklearn import svm

sv = svm.SVC(kernel='linear')

sv.fit(X_train, Y_train)

Y_pred_svm = sv.predict(X_test)

Y_pred_svm.shape

score_svm = round(accuracy_score(Y_pred_svm,Y_test)*100,2)

print("The accuracy score achieved using Linear SVM is: "+str(score_svm)+" %")
```

```
# K Nearest Neighbors

from sklearn.neighbors import KNeighborsClassifier

knn = KNeighborsClassifier(n_neighbors=7)

knn.fit(X_train,Y_train)

Y_pred_knn=knn.predict(X_test)

Y_pred_knn.shape

score_knn = round(accuracy_score(Y_pred_knn,Y_test)*100,2)

print("The accuracy score achieved using KNN is: "+str(score_knn)+" %")


# Decision Tree

from sklearn.tree import DecisionTreeClassifier

max_accuracy = 0

for x in range(200):

    dt = DecisionTreeClassifier(random_state=x)

    dt.fit(X_train,Y_train)

    Y_pred_dt = dt.predict(X_test)

    current_accuracy = round(accuracy_score(Y_pred_dt,Y_test)*100,2)

    if(current_accuracy>max_accuracy):

        max_accuracy = current_accuracy

        best_x = x

#print(max_accuracy)

#print(best_x)

dt = DecisionTreeClassifier(random_state=best_x)
```

```
dt.fit(X_train,Y_train)
Y_pred_dt = dt.predict(X_test)
print(Y_pred_dt.shape)
score_dt = round(accuracy_score(Y_pred_dt,Y_test)*100,2)
print("The accuracy score achieved using Decision Tree is: "+str(score_dt)+" %")

# Random Forest
from sklearn.ensemble import RandomForestClassifier
max_accuracy = 0
for x in range(2000):
    rf = RandomForestClassifier(random_state=x)
    rf.fit(X_train,Y_train)
    Y_pred_rf = rf.predict(X_test)
    current_accuracy = round(accuracy_score(Y_pred_rf,Y_test)*100,2)
    if(current_accuracy>max_accuracy):
        max_accuracy = current_accuracy
        best_x = x
#print(max_accuracy)
#print(best_x)
rf = RandomForestClassifier(random_state=best_x)
rf.fit(X_train,Y_train)
Y_pred_rf = rf.predict(X_test)
Y_pred_rf.shape
```

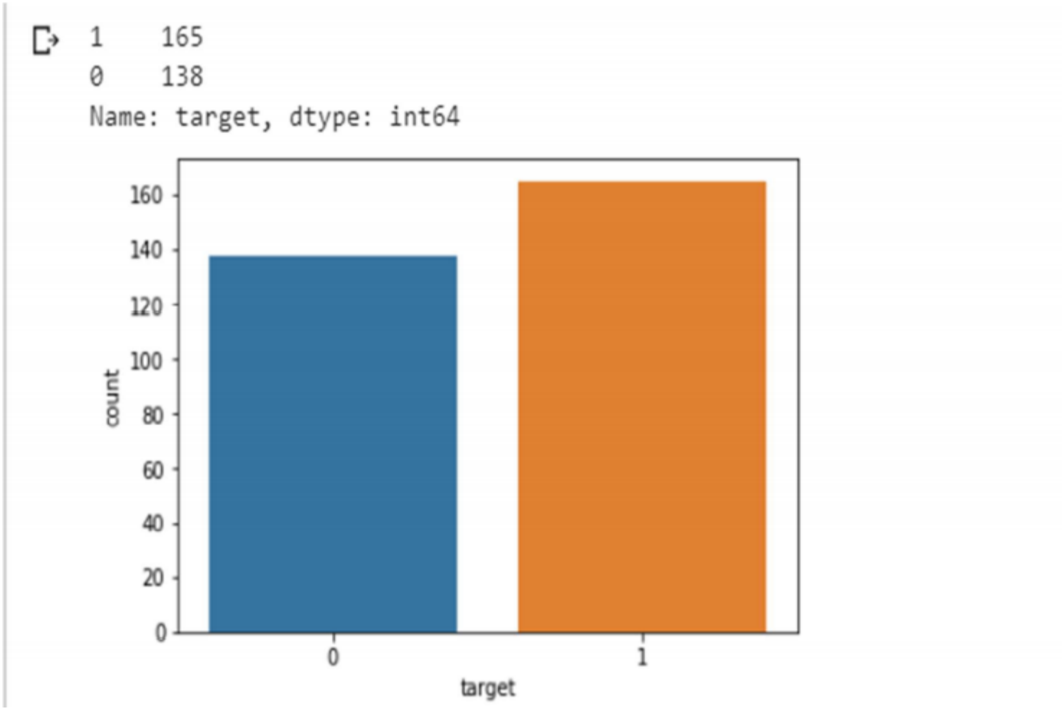
```
score_rf = round(accuracy_score(Y_pred_rf,Y_test)*100,2)

print("The accuracy score achieved using Decision Tree is: "+str(score_rf)+" %")

# VI. Output final score

scores = [score_nb,score_svm,score_knn,score_dt,score_rf]

algorithms = ["Naive Bayes","Support Vector Machine","K-Nearest Neighbors","Decision Tree","Random Forest"]

for i in range(len(algorithms)):

    print("The accuracy score achieved using "+algorithms[i]+" is: "+str(scores[i])+" %")

sns.set(rc={'figure.figsize':(15,8)})

plt.xlabel("Algorithms")

plt.ylabel("Accuracy score")

sns.barplot(algorithms,scores)
```

CHAPTER -6

OUTPUT SCREENS

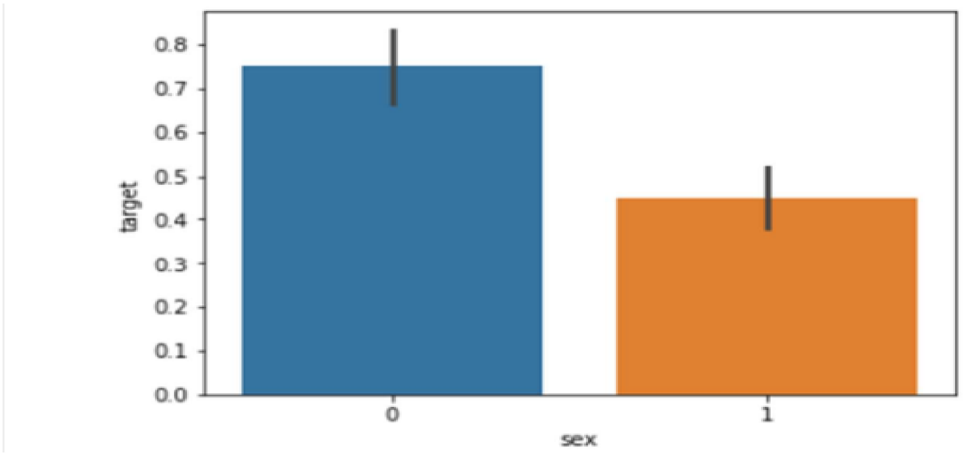
6.1.screen- Analysing target Feature

It shows 0-Disease 1-Non-Disease

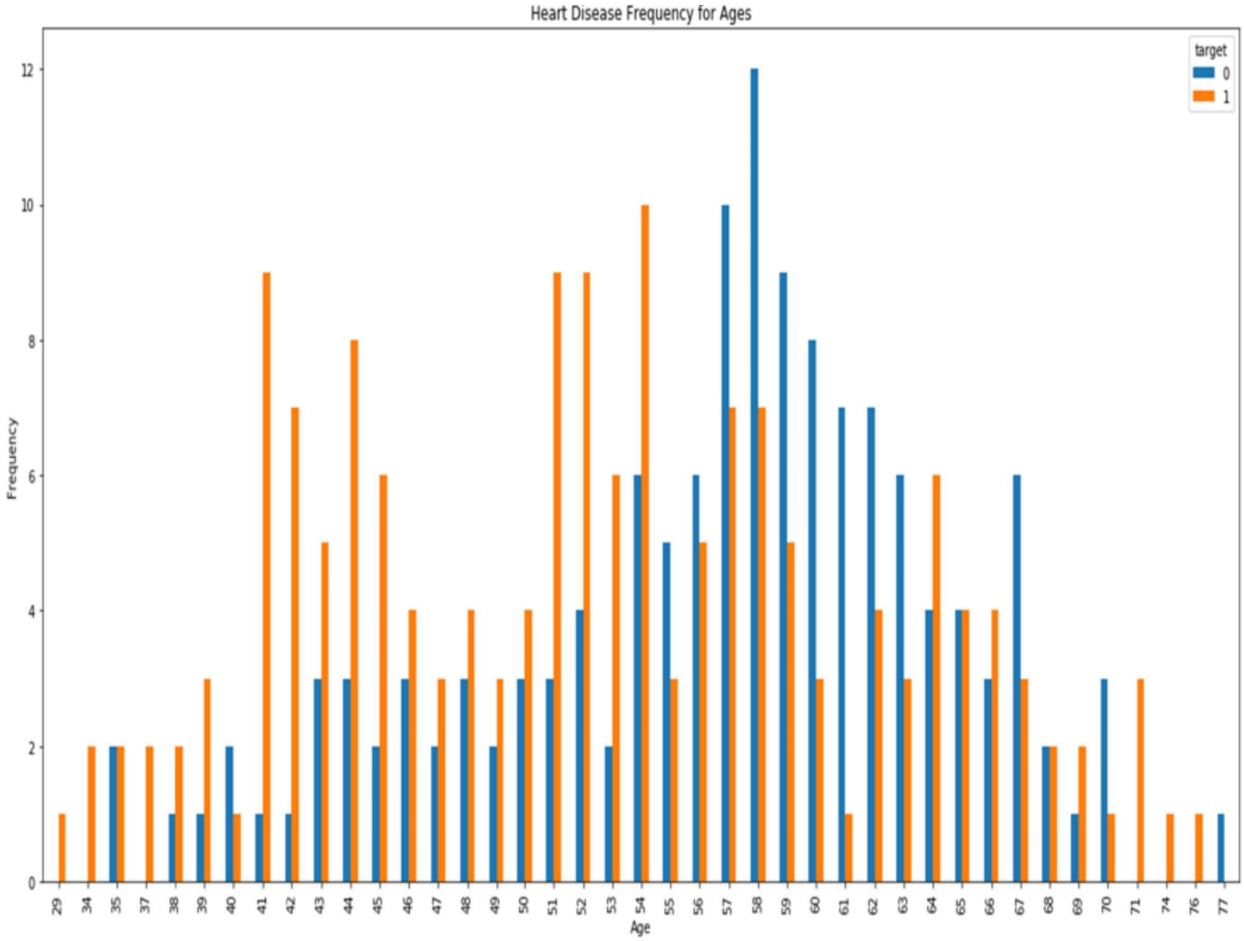


```
1    165
0    138
Name: target, dtype: int64
```

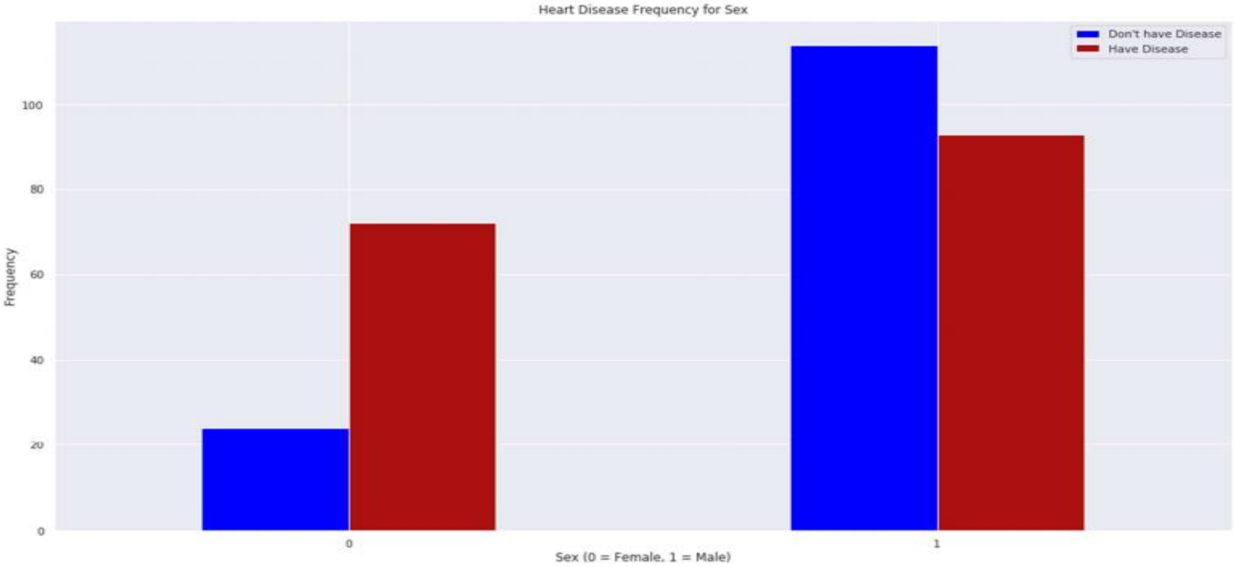
6.2.screen- Analysing Sex Feature



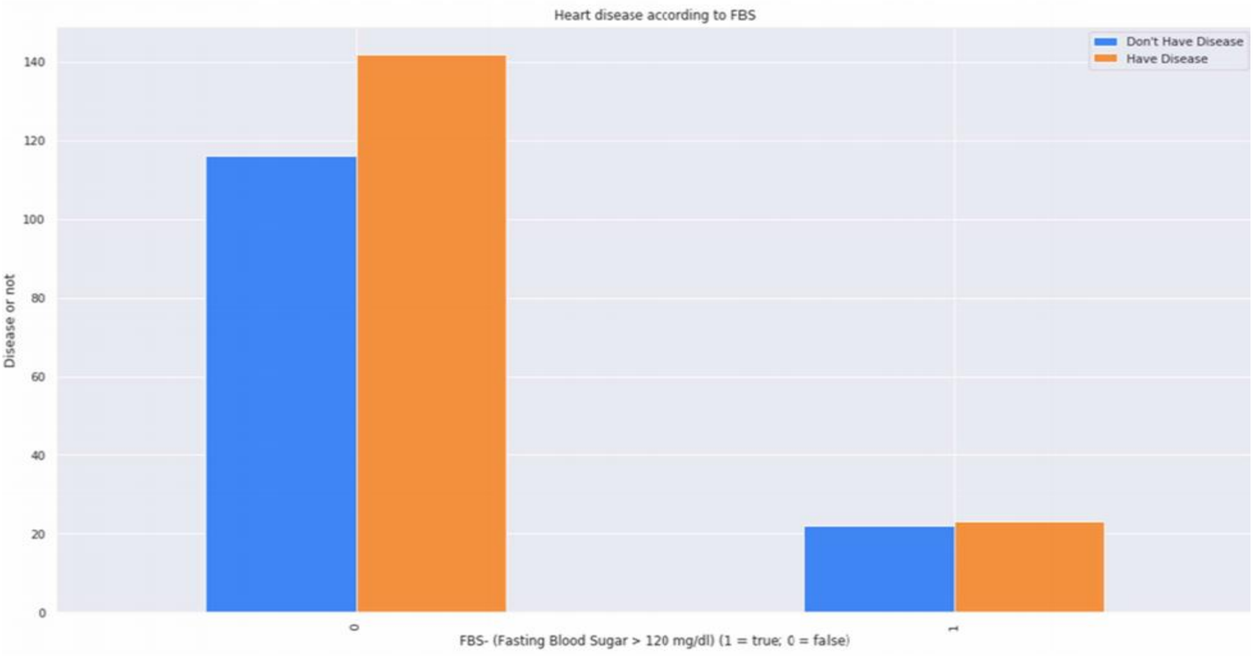
6.3.Screen- Analysing Heart Disease Frequency for Ages



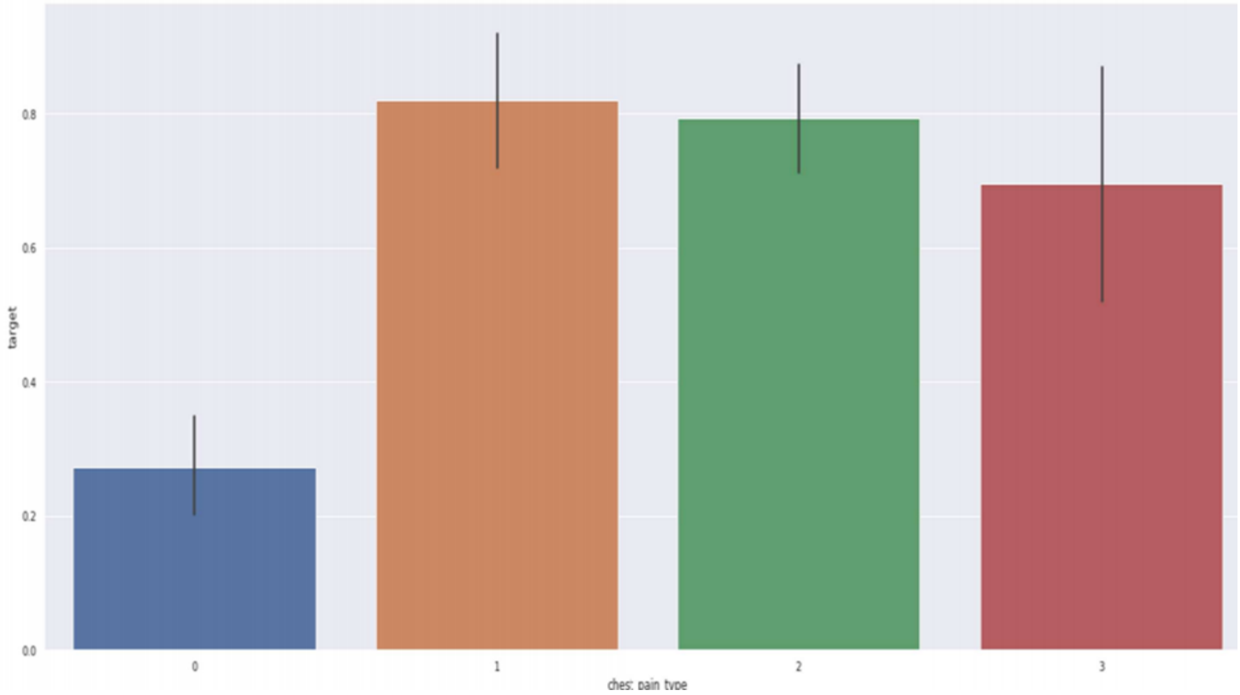
6.4 Screen- Analysing Heart Disease Frequency for Sex



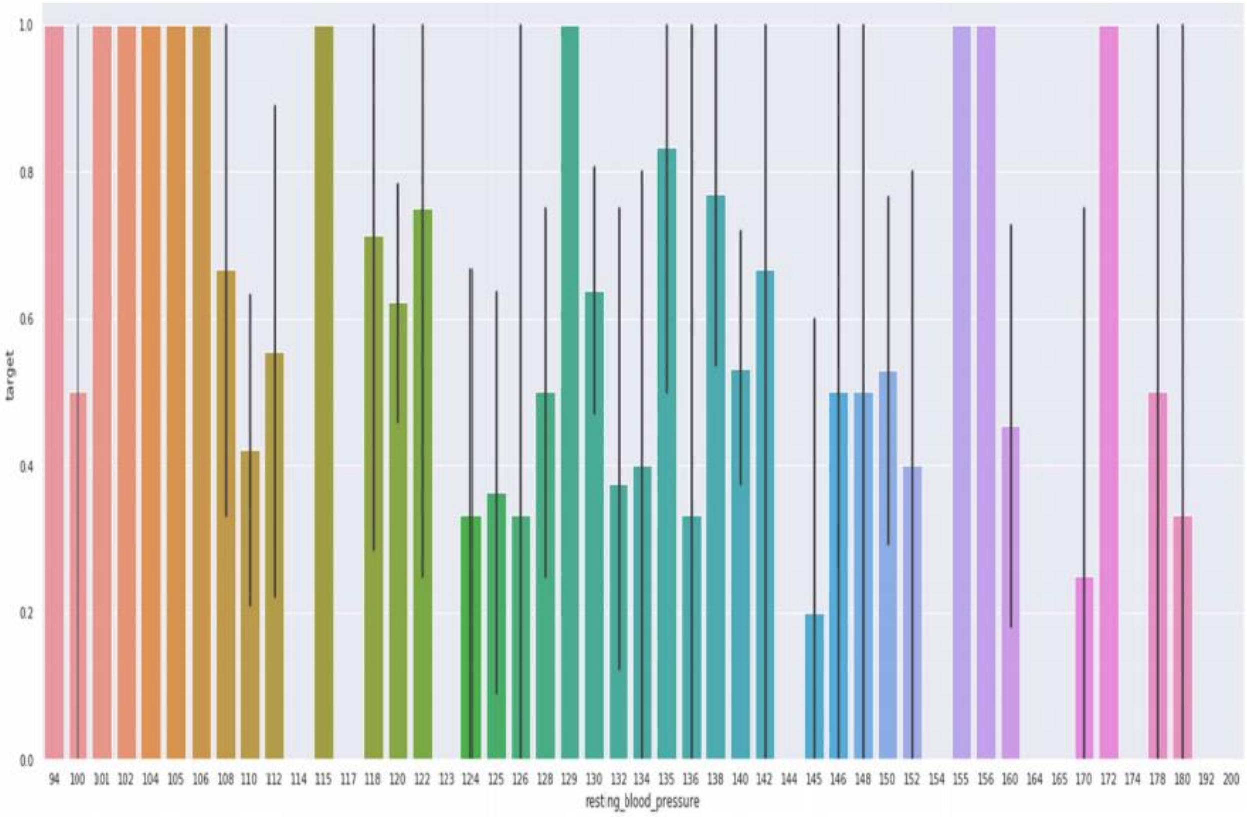
6.5.Screen- Analysis Heart Disease according to Fast Blood Sugar



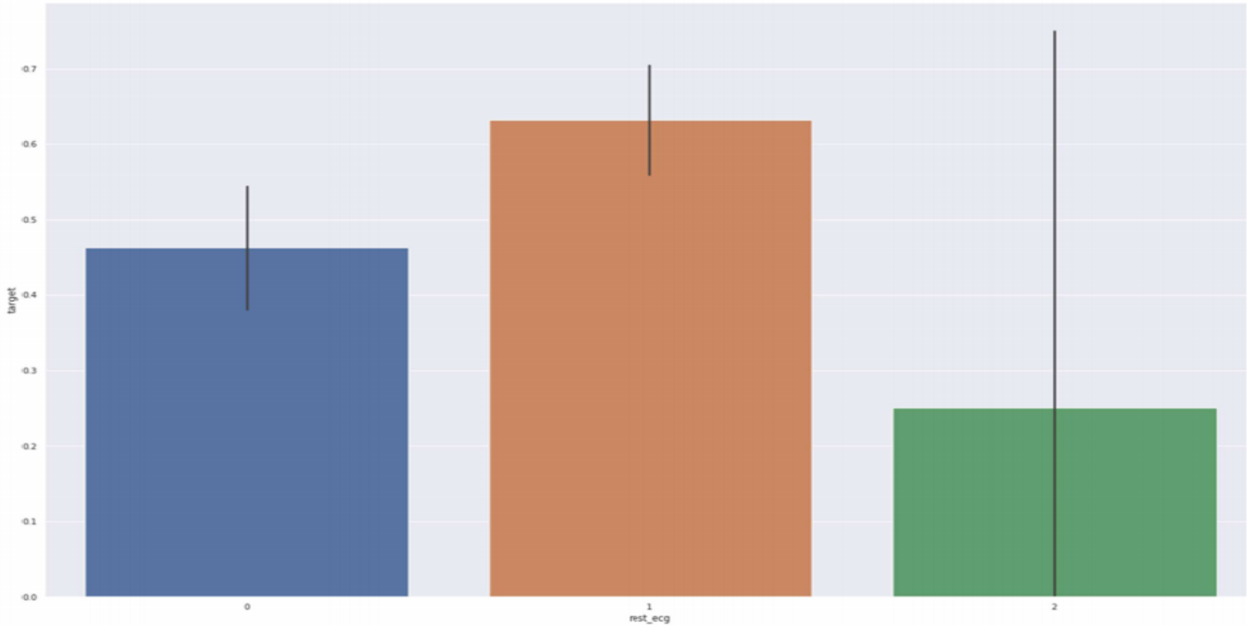
6.6.Screen - Analysis of chest pain type



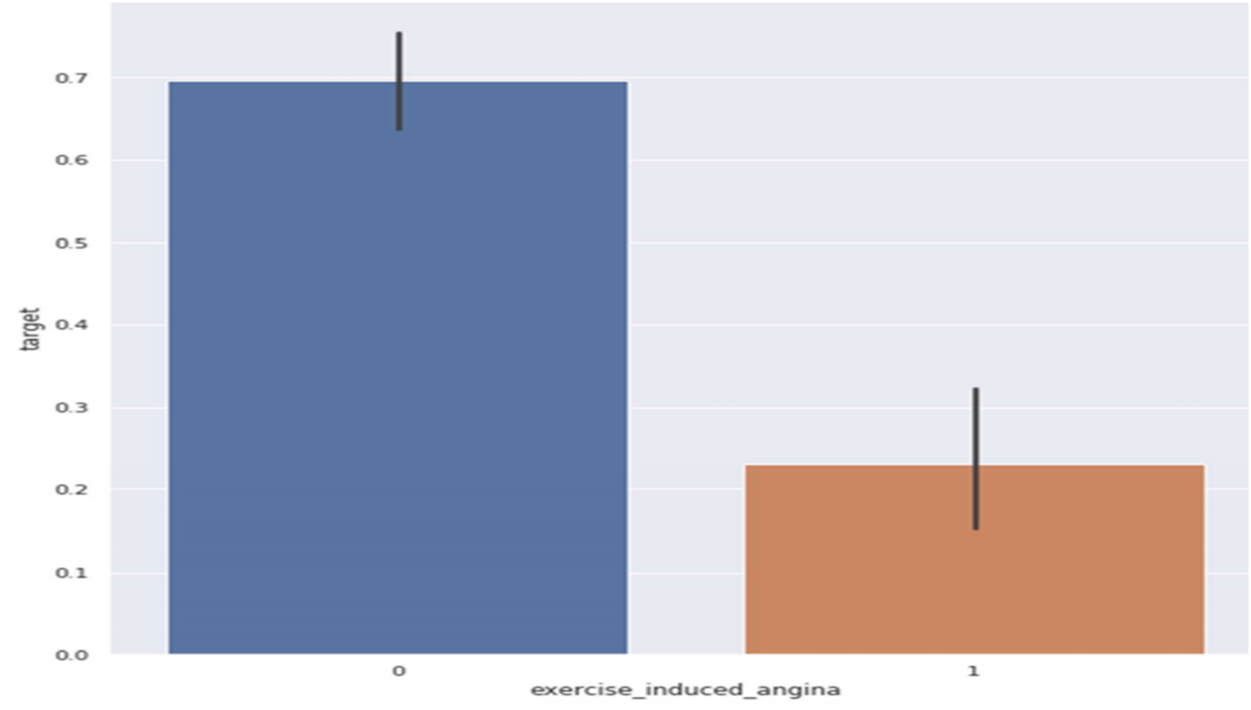
6.7.Screen – Analysing Resting Blood Pressure



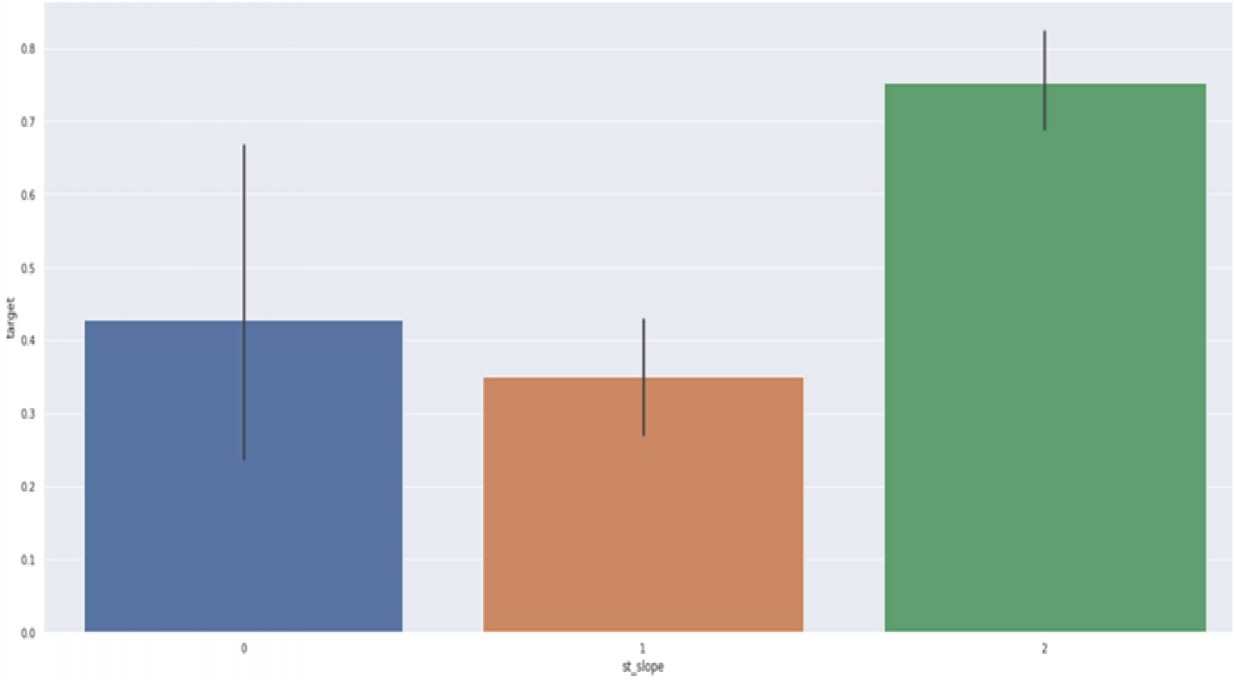
6.8.Screen – Analysing ECG



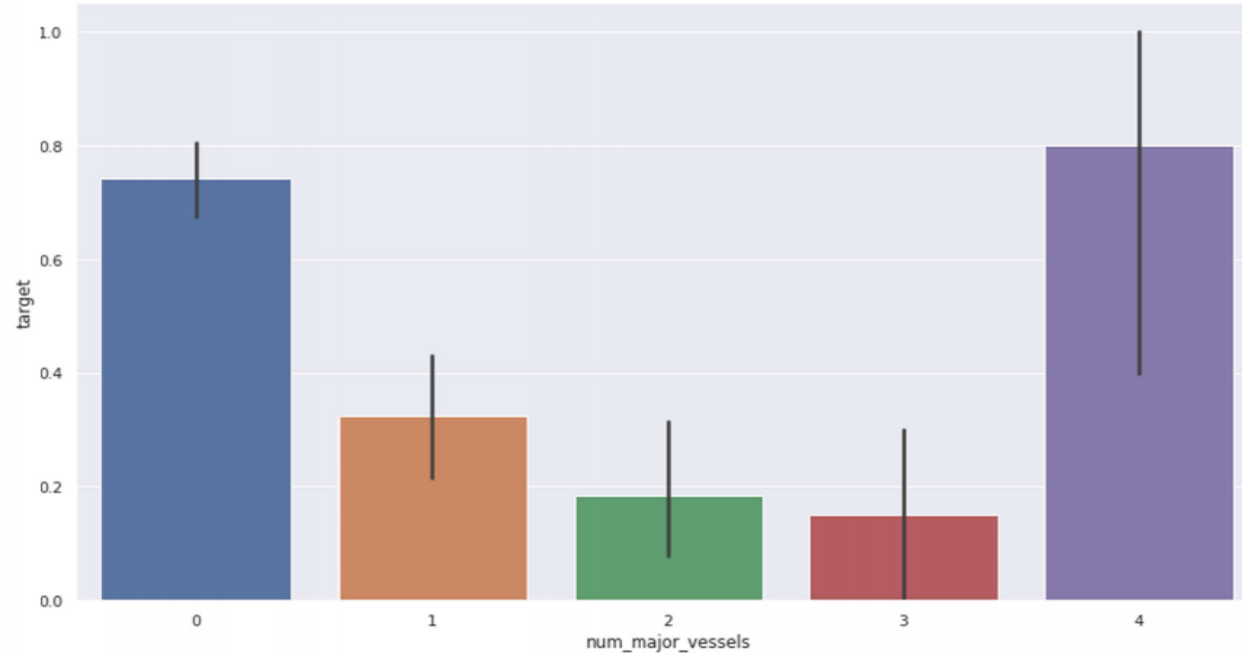
6.9.Screen - Analysing Exercise Incuced Angina



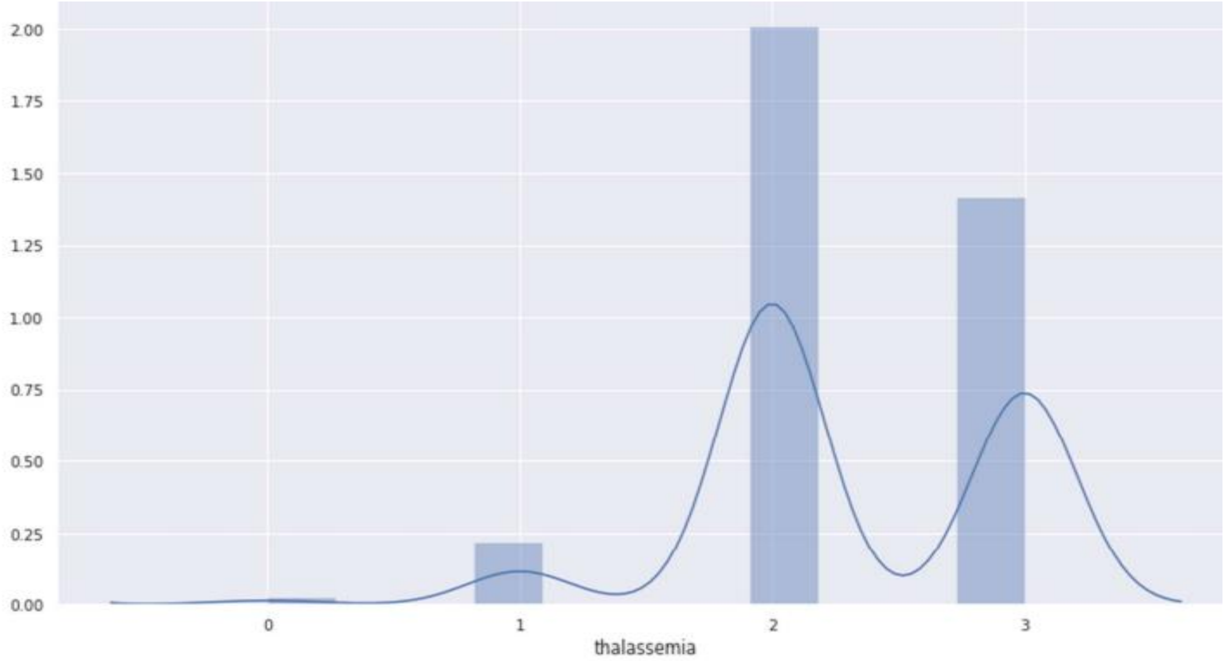
6.10.Screen - Analzing st-slope



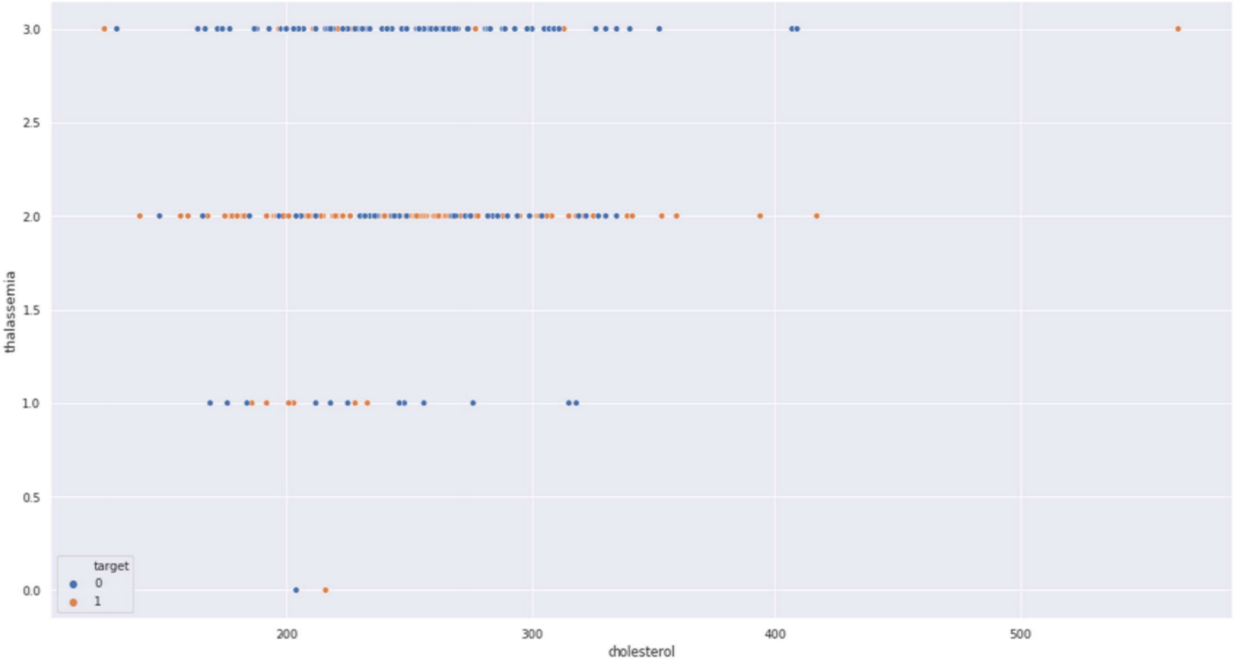
6.11.Screen - Analyzing No of Major Vessels



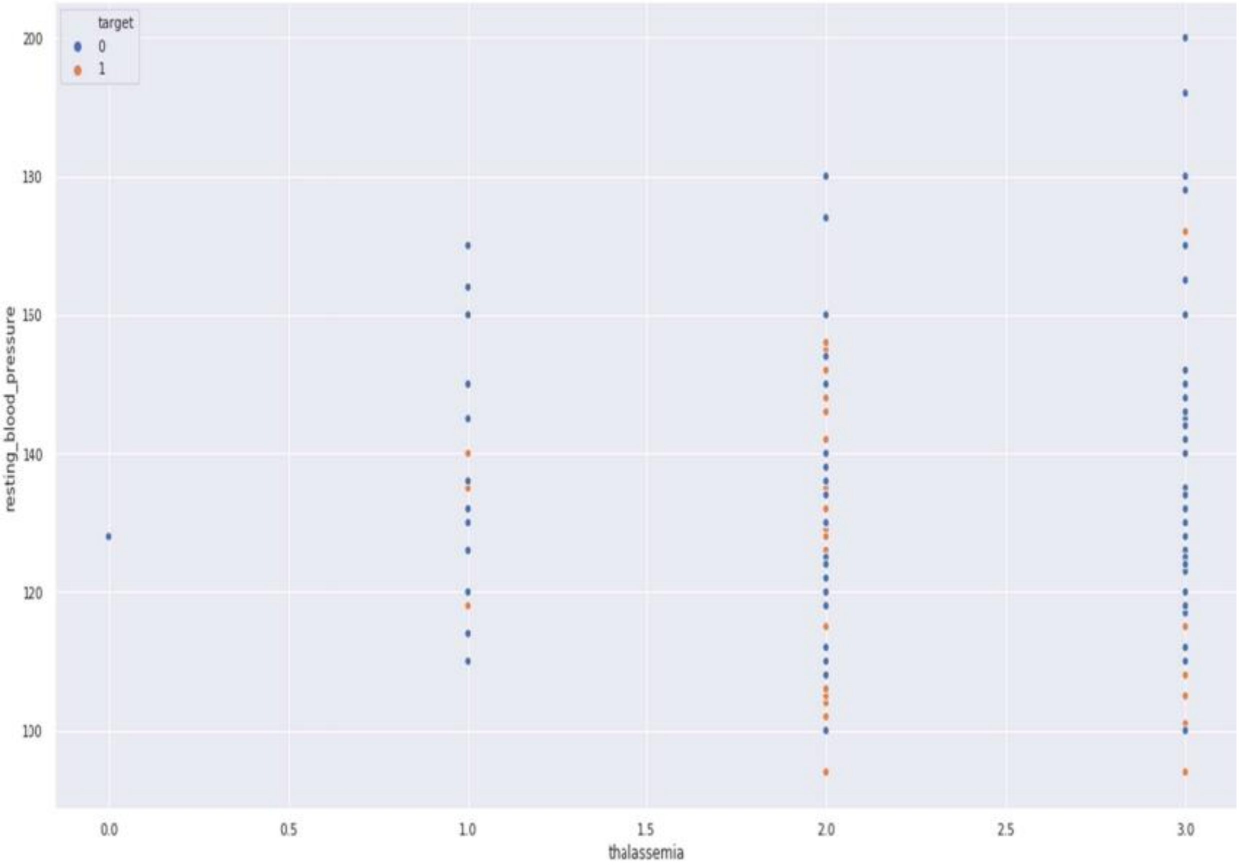
6.12.Screen-Thalassemia Distribution



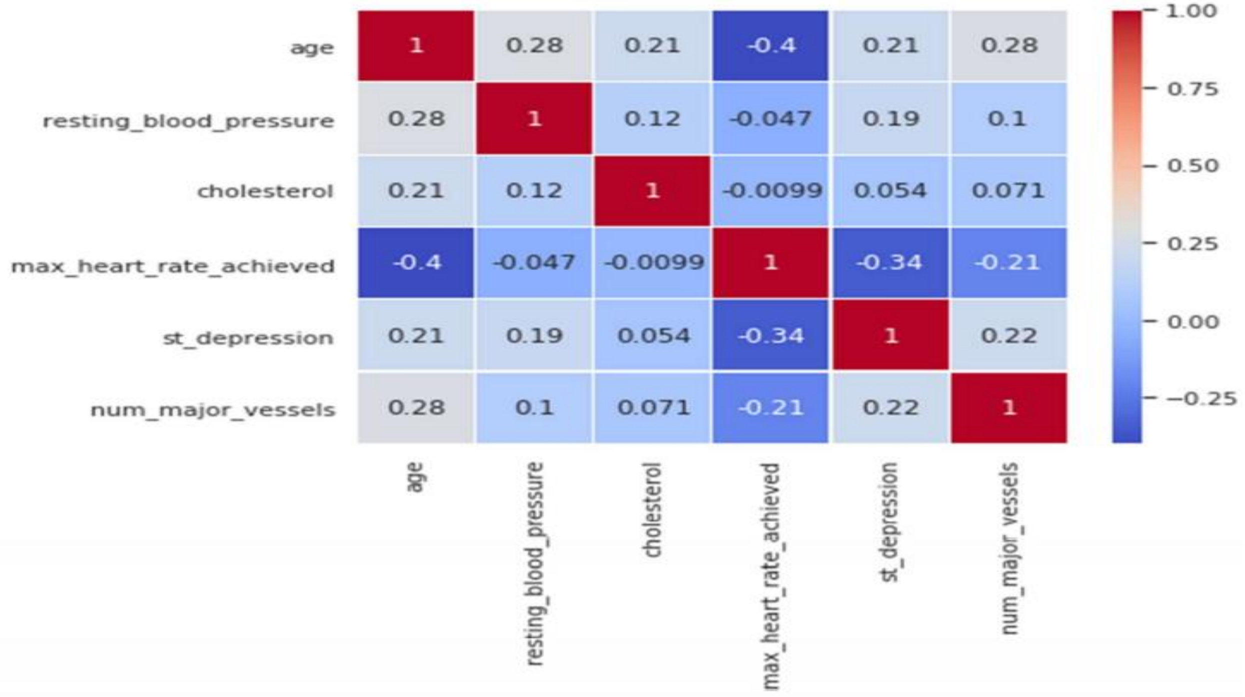
6.13.Screen – Thalassemia vs Cholestrol



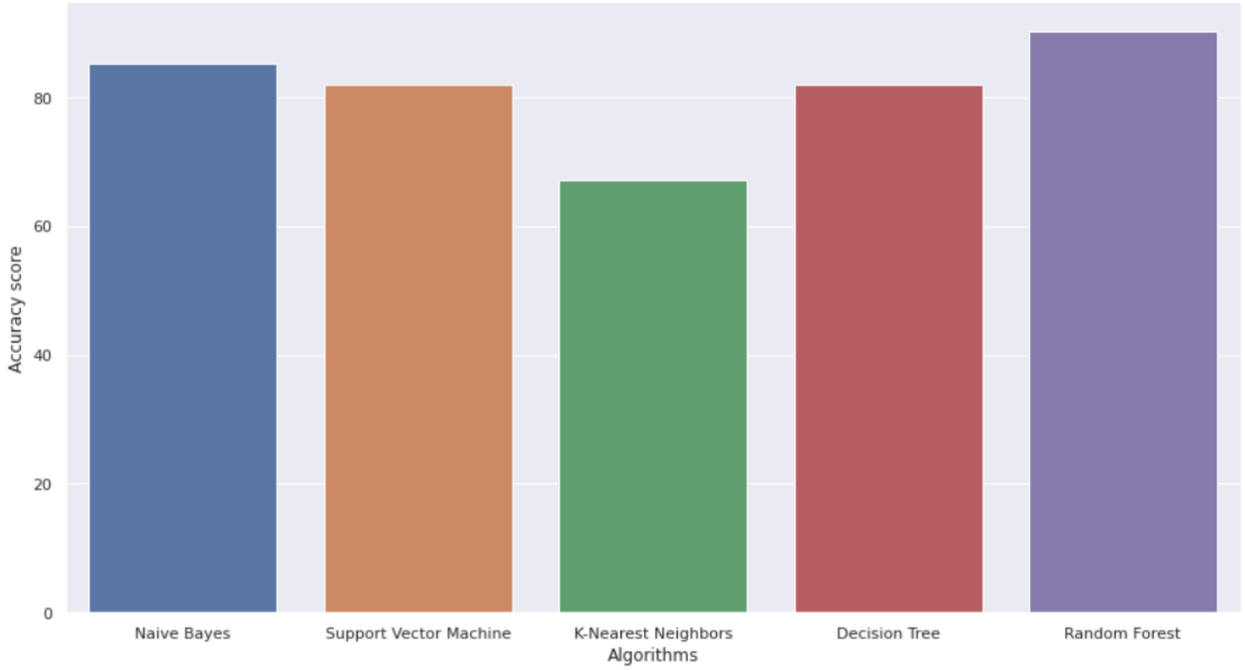
6.14.Screen- Thalassemia vs Resting Blood Pressure



6.15.Screen – Correlation Matrix



6.16.Accuracy score of all algorithms



CHAPTER-7

SYSTEM TESTING

Software testing is an investigation conducted to provide stakeholders with information about the quality of the product or service under test. Software Testing also provides an objective, independent view of the software to allow the business to appreciate and understand the risks at implementation of the software. Test techniques include, but are not limited to, the process of executing a program or application with the intent of finding software bugs.

Software Testing can also be stated as the process of validating and verifying that a software program/application/product:

- Meets the business and technical requirements that guided its design and Development.
- Works as expected and can be implemented with the same characteristics.

7.1 TESTING METHODS

- **Functional Testing**

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

- Functions: Identified functions must be exercised.
- Output: Identified classes of software outputs must be exercised.
- Systems/Procedures: system should work properly

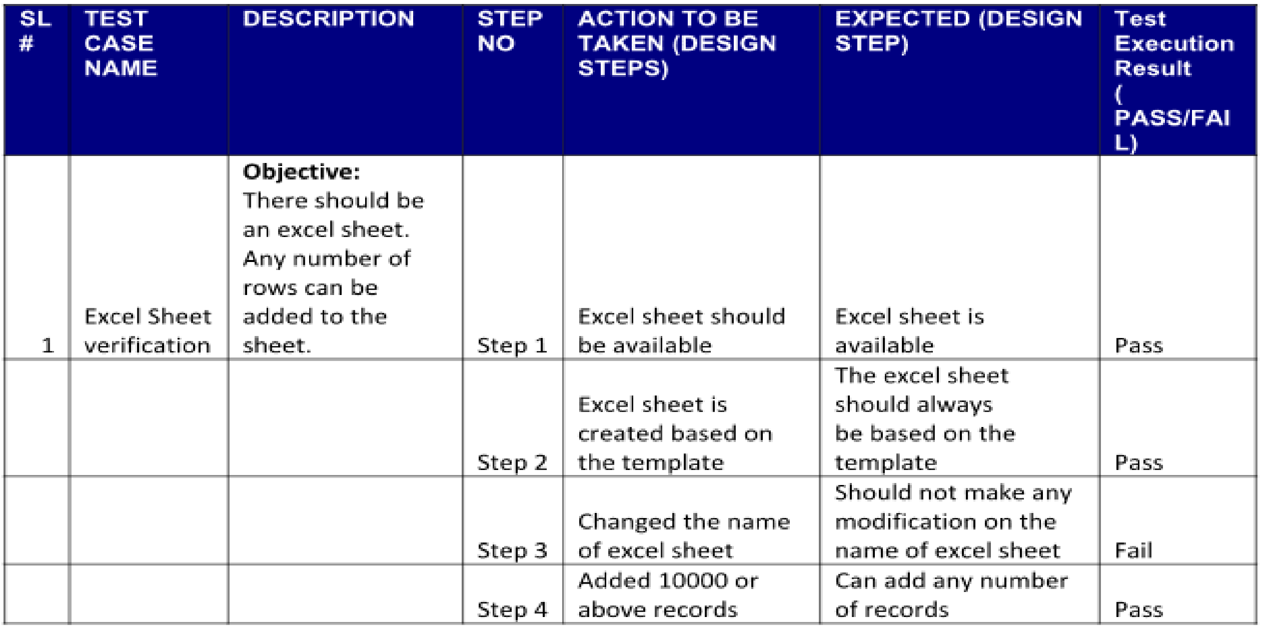
- **Integration Testing**

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

Test Case for Excel Sheet Verification:

Here in machine learning we are dealing with dataset which is in excel sheet format so if any test case we need means we need to check excel file. Later on classification will work on the respective columns of dataset .

Test Case 1 :



SL #	TEST CASE NAME	DESCRIPTION	STEP NO	ACTION TO BE TAKEN (DESIGN STEPS)	EXPECTED (DESIGN STEP)	Test Execution Result (PASS/FAIL)
1	Excel Sheet verification	**Objective:** There should be an excel sheet. Any number of rows can be added to the sheet.	Step 1	Excel sheet should be available	Excel sheet is available	Pass
			Step 2	Excel sheet is created based on the template	The excel sheet should always be based on the template	Pass
			Step 3	Changed the name of excel sheet	Should not make any modification on the name of excel sheet	Fail
			Step 4	Added 10000 or above records	Can add any number of records	Pass

Results

Data mining is a process to extract knowledge from existing data. It is used as a tool in banking and finance, in general, to discover useful information from the operational and historical data to enable better decision-making. It is an interdisciplinary field, the confluence of Statistics, Database technology, Information science, Machine learning, and Visualization. It involves steps that include data selection, data integration, data transformation, data mining, pattern evaluation, knowledge presentation. Banks use data mining in various application areas like marketing, fraud detection, risk management, money laundering detection and investment banking.

CHAPTER-8

CONCLUSION

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 & SVM. 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 SVM, 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

FUTURE ENHANCEMENT

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

CHAPTER-9

BIBLOGRAPHY

[1] Mohan, Senthilkumar & Thirumalai, Chandra Segar & Srivastava, Gautam. (2019). Effective Heart Disease Prediction using Hybrid Machine Learning Techniques. IEEE Access. PP. 1-1. 10.1109/ACCESS.2019.2923707.

[2] A. N. Repaka, S. D. Ravikanti and R. G. Franklin, "Design And Implementing Heart Disease Prediction Using Naives Bayesian," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2019, pp. 292-297.doi: 10.1109/ICOEI.2019.8862604.

[3] A. Ed-Daoudy and K. Maalmi, "Real-time machine learning for early detection of heart disease using big data approach," 2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), Fez, Morocco, 2019, pp. 1-5. doi: 10.1109/WITS.2019.8723839.

[4] Amin Ul Haq, J. P.Li ,M.H.Memon,Shah Nazir and Ruinan Sun," A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms", Wearable Technology and Mobile Applications for Healthcare, Volume 2018 |Article ID 3860146 | 21 pages | https://doi.org/10.1155/2018/3860146.

[5] M. S. Satu, F. Tasnim, T. Akter and S. Halder, "Exploring Significant Heart Disease Factors based on Semi Supervised Learning Algorithms," 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), Rajshahi, 2018, pp. 1- 4.doi: 10.1109/IC4ME2.2018.8465642.

[6] R. Sharmila, S. Chellammal, "A conceptual method to enhance the prediction of heart diseases using the data techniques", International Journal of Computer Science and Engineering, May 2018.

[7] Mr. Chala Beyene, Prof. Pooja Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques", International Journal of Pure and Applied Mathematics, 2018.