A Project Report

on

CARDIOVASCULAR RISK ASSESSMENT USING MACHINE LEARNING ALGORITHMS AND NEURAL NETWORKS

carried out as part of the ***internship/major project*** *Submitted*

by

ZEEL TANNA
209309048

in partial fulfilment for the award of the degree of

Bachelor of Technology

in

Data Science and Engineering

MANIPAL UNIVERSITY
JAIPUR

INSPIRED BY LIFE

School of Information Technology
Department of Data Science and Engineering

MANIPAL UNIVERSITY JAIPUR
RAJASTHAN, INDIA
2024

Manipal University Jaipur

MANIPAL
UNIVERSITY JAIPUR
(University under Section 2(f) of the UGC Act)
INSPIRED BY LIFE

Date: 10/07/2024

CERTIFICATE

This is to certify that the project titled **Cardiovascular Risk Assessment using Machine Learning algorithms and Neural Networks** is a record of the bonafide work done by **Zeel Tanna (209309048)** submitted in partial fulfilment of the requirements for the award of the Degree of Bachelor of Technology (B.Tech) in **Data Science & Engineering of Manipal University Jaipur, during the academic year 2023-24.**

Dr. Dinesh Sharma

Project Guide, Dept of Data Science

Manipal University Jaipur

Dr. Akhilesh Kumar Sharma

HOD, Dept of Data Science

Manipal University Jaipur

ABSTRACT

India is often referred to as “heart disease capital of the world” due to the high prevalence of cardiovascular diseases and the significant impact these diseases have made on the population. Various scientific studies have observed that Indians are in general at a higher risk of developing heart diseases at a younger age compared to population in Western Countries.

Factors such as genetic predisposition, high rates of diabetes, hypertension, obesity and lifestyle changes associated with urbanization and development are few of the many observed factors that contribute to increment in rate of being diagnosed with possible Heart Disease Risk. Additionally, challenges related to early detection, awareness and healthcare access also affect the likelihood of contracting Heart Disease. Statistics show that about 16% of the overall deaths globally is only due to heart diseases. Around 532 million people are suffering from heart diseases or have been diagnosed as a patient of Heart related diseases.

This project aims at identifying individuals who are at a higher risk of experiencing a heart attack in the immediate future or later. The main focus would be to analyze various risk factors such as age, gender, medical history, lifestyle factors and biometric measurements to make certain predictions regarding the condition of the patients or users.

By utilizing various available advanced tools and technology such as Python, CSS, Html, and JavaScript I intend to develop a prototype web page such that the user can self-diagnose whether he or she is a potential patient for Heart Disease or not. By implementing various available advanced Machine Learning models, a trained model will be selected with a high range of accuracy and great robustness to various data, which will analyze the user input in the background and display prediction results on the webpage itself.

Inculcating the use of advanced technology will not only benefit the Health Care industry in the present as well as in the future. By implementing techniques like Artificial Intelligence, will help automate several processes like Treatment Suggestions, Rapid Diagnosis, In-Depth analysis of Disease afflicted to the patient, Visual Representation of Diagnostics and suggestions for consulting the doctor with well expertise in the similar field.

LIST OF FIGURES

Table of Contents

1. Introduction

1.1 Introduction

Heart Disease remains as one of the leading causes of morbidity and mortality worldwide. Early and accurate prediction of heart disease can significantly improve the patients chances of prevention or prediction by enabling timely interventions and personalized treatment plans. Traditional methods of diagnosing heart disease often relied on the clinical expertise of the doctor diagnosing the reports and a series of diagnostic tests which can be very time-consuming and exhausting. With the inculcation of Technology in the medical field, diagnosing and formulation of treatment plans have advanced by leaps and bounds. Machine learning has been providing powerful tools to analyze large datasets and uncover complex patterns that may not be immediately apparent to human practitioners.

This report explores the application of machine learning techniques to predict heart disease. By leveraging a variety of algorithms and data from patient records, I aim to develop predictive models that can assist healthcare professionals in identifying individuals that are at high risk of developing heart disease.

The primary objective of this study is to evaluate and build different models that can be practically used in a clinical setting. For us to accurately assess the possibility of a disease a single parameter or few parameters may prove to be ineffective. Hence, I decided to combine a total of 13 features and a total of 303 records of values to examine and improve the accuracy of various models such as Logistic Regression, K Nearest Neighbors, Random Forest, Support Vector Machine, Decision Trees and many more.

By integrating machine learning into the diagnostic process, we hope to contribute to the development of more efficient and accurate tools for heart disease prediction, ultimately improving patient care and outcomes.

1.2 Problem Statement

Despite advancements in medical science, early diagnosis and intervention remain a critical challenge due to the multifaceted nature of diseases and the variability in symptoms observed. Traditional methods, though effective, often require extensive testing and can be time consuming, which is of utmost importance with respect to a patients condition and future developments.

The primary challenge is to develop an accurate, efficient and scalable method for predicting heart disease risk that can be easily integrated into routine clinical practice. This method should utilize available patient data, such as demographic information, medical history, clinical measurements and lifestyle factors. The specific problem to address is : How can machine learning models be effectively utilized, so that we can enhance early diagnosis and timely start medicinal interventions.

1.3 Objectives

The main objectives of this extensive project are to:

- Gathering a robust dataset that includes relevant features associated with heart disease

- Identifying suitable machine learning models and training them on the dataset to develop predictive models

- Assessing the performance of the models using appropriate metrics to ensure integrity, specificity and accuracy.

- Determining the most significant predictors of heart disease and their percentage of influence in heart disease predictions.

- Develop a user-friendly tool that can be used by healthcare professionals as well as normal users to predict heart disease risk in real time clinical settings.

1.4 Scope of Project

The scope of Heart Disease Prediction can be implemented in various fields such as enhanced healthcare delivery, personalized preventive care, routine health assessments, optimize resource utilization, drug developments, health tracking and various other fields.

By extending the use of Heart Disease Prediction tools in various domains we can create a comprehensive approach at managing and preventing heart diseases, ultimately leading to a better and health life.

2. Background Detail

2.1 Conceptual Overview

Conceptual Overview includes the understanding of all the process of Machine Learning modelling and analysis, some of the terminology used is explained below:

The foundation for any machine learning model is data. For this project, relevant data includes:

- Demographic Information: Age and Gender
- Medical History: Family history of Disease and previous medical conditions
- Clinical Measurements: Blood Pressure, Cholesterol levels, Blood Sugar Levels
- Lifestyle Factors: Smoking Habits, Physical Activity, diet, alcohol consumption
- Diagnostic Tests: ECG results, echocardiograms, stress tests

Before data is fed into machine learning models it undergoes several steps of preprocessing which includes data cleaning, feature engineering, normalization or standardization and encoding of categorical variables.

During the data cleaning process , statistical and mathematical analysis of data is done such as determining the maximum value, count of null values, count of error values, minimum value and many more parameters. Then errors and null values are rectified according to requirements of the model such that they don’t affect the results and accuracy of the models.

As for the feature engineering and the encoding process, various variables that have textual data in them or categorical data in them are converted into numerical representations such that machine learning models can easily identify them for processing. Attributes such as sex, target, and fbs is converted in 0 and 1 form from categorical values such as yes/no and male/female.

Exploratory Data Analysis involves analyzing the dataset to understand the underlying patterns, correlations and distribution of data which helps determine which models are suitable for the dataset being used. As a part of EDA, I conducted analysis of various features and their distribution. Each attributes distribution and count values were represented in the form of graphs using Bar plots and Count plots.

Correlation Analysis is a crucial step of Exploratory Data Analysis. It helps identify relationships between features and the target variable by calculating correlation coefficients. Hence, by constructing a correlation heatmap identifying variables that heavily influence the decision making of target factor was clearly made apparent.

Several Machine Learning algorithms are suitable for determining whether the patients are targets or not, but each model has its own strengths and weaknesses. Since different models provide different accuracies, choosing the one with the highest accuracy is an obvious conclusion we can come to. Thus, I made use of multiple models such as Regression models, Decision Trees, Random Forest, SVM, KNN, Gradient Boost and Neural Networks to identify the one with most accuracy.

The final step of the project includes evaluating the models performance using various metrics such as Accuracy, Precision values, Recall values and then deploying the trained model into real world application. The model that is robust and is highly accurate will not only be beneficial for clinical treatments but also for clinical diagnosis.

2.2 Machine Learning Algorithms

Machine Learning algorithms are computational models that enable computers to interpret patterns and trends and forecast inferences or judgements based on data without the need to involve explicit programming. These algorithms are used in a wide range of applications and each of them is suited for different tasks and data structures. Given below are brief descriptions of all the Machine Learning algorithms that have been utilized in this project.

1. Logistic Regression

Regression algorithm is used to understand the relationship between independent and dependent variables. It is commonly used to make projections such as whether a customer will churn, email is spam or not, or whether a patient has a particular disease based on medical tests. While Linear Regression is leveraged when dependent variables are continuous, logistic regression is chosen when the dependent variable is categorical in nature, meaning there are binary outputs such as “true or false” or “yes or no” or “1 or 0”. Logistic Regression is a supervised learning algorithm that is primarily utilized for binary classification tasks. It assumes a linear relationship between the features and log-odds of outcome. Logistic regression is also crucial when users want to interpret the impact of each feature or attribute on the outcome of the model.

2. K-Nearest Neighbors (KNN)

The K Nearest Neighbors or KNN algorithm is one of the simplest Machine Learning algorithms based on Supervise Learning techniques. It is an instance-based learning algorithm that is utilized for both classification and regression tasks. It predicts the class of a data point based on the majority class of its K-Nearest Neighbors in the feature space. The algorithm utilizes the mathematical concept of Euclidean Distance to calculate the distance between K number of neighbors and assign the data points to that category for which the number of neighbors having similarity is maximum. KNN is mainly used for classification purposes hence it is applied in recommendation systems, pattern recognition, and anomaly detection where the locality of data points play a critical role in prediction.

3. Support Vector Machines (SVM)

It is a supervised machine learning algorithm that excels in both regression as well as classification tasks. SVM can handle both linearly separable as well as non-linearly separable data by using kernel functions like sigmoid, polynomial or RBF function. SVMs function by finding the optimal hyperplane that separates classes in a high dimensional space. The main goal of the algorithm is to identify the hyperplane that has a margin that is closest to the points of different classes. SVM algorithm can be used for a variety of tasks such as text classification, image classification, spam detection, face detection and anomaly detection.

4. Naïve Bayes Classifier

The naïve bayes classifier is a supervised machine learning algorithm that works on the basis of collection of algorithms based on the principle of Bayes Theorem. It is not a single algorithm but a family of multiple algorithms where they all share a common principle that is; every pair of features being classified is independent of each other. The model predicts the probability that an instance belongs to a class with a given set of feature values. The Naïve Bayes theorem is usually used in classification scenarios such as spam filtering, sentiment detection, heart disease prediction and many more problems.

5. Decision Trees

Decision tree is a type of supervised machine learning algorithm that is capable of performing classification as well as regression tasks. The model partitions the data into subsets based on the most significant attributes at each step of the process, forming a tree like structure. The major nodes are referred to as parent nodes and the branches are referred to as child nodes. The parent nodes represent a "decision", and the child nodes or leaf nodes represent the outcome or prediction. Decision trees selects the best attribute to split the data using metrics like Gini index, Entropy or Information Gain. These sub-datasets help the model reach the most optimal solution possible. Decision trees are known to be susceptible to overfitting especially when there a large number of nodes but despite that they are used since they are easy to understand and interpret.

6. Random Forest

The random forest algorithm is an ensemble learning method that is based on the principle of decision tree algorithm. It is designed to improve the performance of decision trees and reduce the overfitting effect of single decision trees. The algorithm works by building multiple decision trees during the training process and outputs the final result by combining results using statistical concepts such as mean or mode. The algorithm is more robust in nature as compared to traditional decision trees and has improved accuracy as well. They are widely utilized in domains such as finance, healthcare and bioinformatics.

7. Extreme Gradient Boosting (XG Boost)

The Extreme Gradient Boosting algorithm is an optimized and powerful gradient boosting algorithm that is known for its efficiency and parallel processing performance. The algorithm is based on the implementation of Gradient Boosting to decision trees. The decision trees are arranged in a sequential form of schema. Weights are assigned to all the independent variables and then are fed into the decision tree that produces the final results. XG Boosting algorithm has a strong record of producing high quality results but is computationally intensive at the same time and is time consuming since multiple hyperparameters need to be tuned to produce optimal results.

8. Neural Networks

Neural Networks models are a part of deep learning algorithms that are inspired by the structure and functioning of the human brain. They consist of multiple interconnected layers of nodes called neurons that process information. Neural Networks are known to excel in adapting to complex patterns and relationships from large datasets, making them highly suitable for tasks such as speech recognition, Natural Language Processing and many other such complex tasks. Meanwhile neural networks are powerful, but they have their downsides as well. For efficient training, neural networks need sizable amounts of datasets otherwise their performance may suffer due to skewed data. They require large amounts of processing power and are susceptible to overfitting.

2.3 Technology Used

Python is a high-level all-purpose programming language that supports multiple programming paradigms including structured, object-oriented and functional programming. Multiple python libraries such as Numpy, Pandas, Seaborn and many more are leveraged to perform data analysis, machine learning modelling and for web development.

The project itself can be divided into 3 major components: Machine learning model and Exploratory Data Analytics, Flask API and the Web Page. The first component of the project involves collecting data and filtering out the relevant features for Exploratory Analysis and ML modelling. Then using libraries such as NumPy we identify and calculate crucial statistical measures such as Mean, Median, Mode, Minimum, Maximum and many more, which help understand and interpret the distribution of data points in the dataset. The Pandas library is used to load, clean and preprocess the data. It facilitates easy manipulation of dataframes and handling of missing values which is crucial for ensuring that the dataset is ready for machine learning models.

To understand and interpret essential characteristics of the preprocessed dataset I utilization of libraries such as matplotlib and seaborn was a must. They helped visualize the data through histograms, bar plots, box plots and correlation matrices. It is a necessary step that helps identify patterns, relationships between datapoints and potential outliers in the dataset. With the preprocessing of data done and identification of crucial patterns accomplished we ultimately reach the most important part of the project, that is, training and testing multiple Machine Learning models with the dataset.

The Scikit Learn library is a crucial library in terms of model building, model training and accuracy predictions. It provides all the necessary functions and measures for the complete process of making predictions using ML models and determining the accuracy of their predictions. The project includes utilizing the in-built ML models to train and test the data that is split into the ratio of 80% is 20% and deduce predictions based on the same. Accuracy metrics such as Precision, Recall, F1 score, and Support are used to determine the model with the highest accuracy.

The model is then saved and connected to a Web Page that is built using languages like CSS, JavaScript and HTML. The model is connected using Python’s inbuilt API called Flask API. The model is made to be utilized in the background of the webpage such that the Web Page takes in user input, the model analyzes and makes predictions, and the Flask API facilitates the complete connection process between the webpage and the machine learning model.

3. Methodology

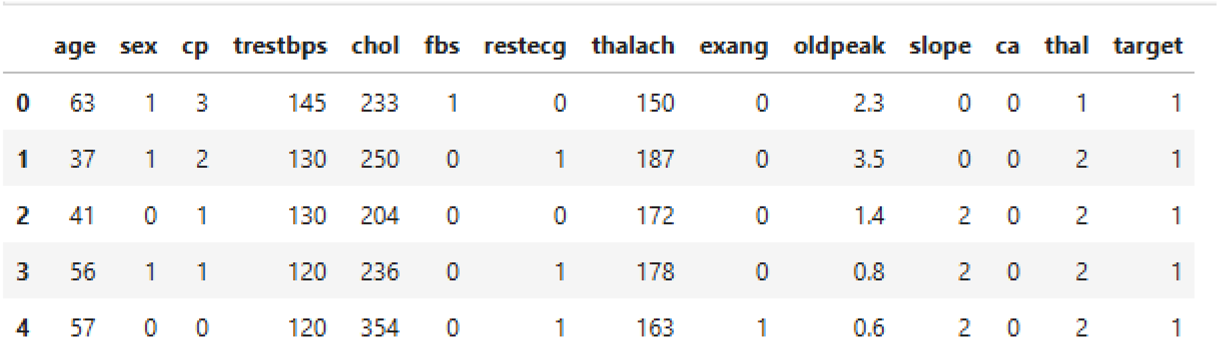
3.1 UCI Heart Disease Dataset

The UC Irvine Heart Dataset is a widely used dataset in the field of medical research and Machine Learning for developing and testing models for Heart Disease Assessment. It includes data from several sources such as Cleveland Clinic Foundation, the Hungarian Institute of Cardiology, and the University Hospital in Zurich.

The dataset consists of various patient attributes that are potentially predictive of heart disease. The target variable indicates the presence or absence of heart disease in the patient. The 14 attributes that are included in the dataset are as follows:

1. **Age:** Age of Patients in years
2. **Sex:** Gender of the Patient (1 = male , 0 = female)
3. **Chest Pain type (4 values – Ordinal):**
    a. Value 1: typical angina
    b. Value 2: atypical angina
    c. Value 3: non-anginal pain
    d. Value 4: asymptomatic.
4. **(trestbps) resting blood pressure:** Resting blood pressure in mm Hg on admission to the hospital
5. **Serum Cholesterol (Chol)**: Serum cholesterol in mg/dl
6. **Fasting Blood Sugar (fbs)**: Fasting blood sugar > 120 mg/dl (1 = true, 0 = false)
7. **Resting Electrocardiographic Results (restecg)**:
    a. Value 0: Normal
    b. Value 1: Having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
    c. Value 2: Showing probable or definite left ventricular hypertrophy by Estes' criteria.
8. **Maximum Heart Rate Achieved (Thalac)**: Maximum heart rate achieved during the test
9. **Exercise Induced Angina (exang)**: Exercise-induced angina (1 = yes, 0 = no)
10. **ST Depression Induced by Exercise Relative to Rest (oldpeak)**: ST depression induced by exercise relative to rest
11. **Slope of the Peak Exercise ST Segment (slope)**:
    a. Value 0: Upsloping
    b. Value 1: Flat
    c. Value 2: Downsloping
12. **Number of Major Vessels Colored by Fluoroscopy (ca)**: Number of major vessels (0-3) colored by fluoroscopy

13. **Thalassemia (thal):**
    a. Value 3: Normal
    b. Value 6: Fixed defect
    c. Value 7: Reversible defect
14. **Target Variable (num):** Diagnosis of heart disease (angiographic disease status)
    a. Value 0: < 50% diameter narrowing (no presence of heart disease)
    b. Value 1: > 50% diameter narrowing (increasing severity of heart disease)



	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

Figure 3.1.1 Attributes

Features in Detail:

1. Chest Pain Type: When diagnosing for potential heart issues, medical professionals usually associate chest pain or also known as angina. Angina can be classified into four categories based upon their characteristics and symptoms:

    a. Typical Angina: Occurs when over-exertion or emotional stress is involved and is resolved with rest or medication. Typical angina is usually strongly associated with Coronary Artery Disease and indicates higher risk of Heart Disease.
    b. Atypical Angina: Chest pain that does not fit the criteria of angina but is yet considered of being suggestive of heart related issues. The amount of risk associated with Heart Disease is weak in comparison to Typical Angina.
    c. Non – Anginal Pain: Chest pain that is neither related to heart nor to coronary artery disease. Although it is not frequently associated with presence of Heart Disease, a thorough evaluation of the condition is still suggested.
    d. Asymptomatic: Although there may be no chest pain involved, the patient may have several other risk factors indicating presence of heart disease.

2. Resting Blood Pressure: Resting blood sugar plays a crucial role in assessing and predicting the risk of heart disease. It measures the pressure exerted by blood against the wall of arteries when the body is at rest. Elevated levels of resting blood pressure indicates the potential existence of Heart Disease.

3. Serum Cholesterol: Cholesterol can be categorized into two different categories: Low Density Lipoproteins (LDL) and High Density Lipoprotein (HDL). High levels of LDL cholesterol indicates that there is a major risk factor for coronary heart disease or heart disease in general. Low levels of LDL and High levels of HDL indicate that the level of risk of heart disease is comparatively less.

4. Fasting Blood Sugar: FBS specifically measures the blood glucose levels after fasting for a certain period of time, which usually is overnight. Elevated fasting blood sugar levels indicate insulin resistance of diabetes which are associated with increased risk of cardiovascular diseases.

5. Resting ECG: Resting Electrocardiogram is a widely known and used tool in diagnosing presence of Heart related issues. ECG usually detects various types of irregular heart rhythms which indicate underlying heart conditions and risk of complications like heart stroke or failure.

6. Thalach: The term thalach refers to the maximum heart rate achieved during stress test or exercise tolerance test. Usually patients with a higher thalach value are considered to be healthy. Whereas patients that have lower values of thalach are considered to be potential heart disease patients.

7. Old Peak: During the exercise stress testing, old peak is quantified as a measure of degree of ischemia. Higher the degree of depression observed in ST depression more is the severe impairment of blood flow, suggesting higher likelihood of heart disease in the patient. An oldpeak having ST depression > 1mm is considered to be abnormal.

8. Slope: The slope of peak exercise ST segment refers to the ECG measurement taken during the peak duration of exercise induced stress test. During the stress test if the slope of the peak exercise ST segment is positive or up-sloping, it indicates the patient is health and shows signs of gradual increase in elevation. Whereas a patient that shows signs of horizontal or down-sloping ST segment during peak exercise stress indicates abnormality and suggests the presence of heart disease or in particular myocardial ischemia.

9. Coronary Artery Vessels (ca): Number of Major vessels colored by fluoroscopy refers to the number of major coronary arteries that show significant amount of narrowing during coronary angiography or guided imaging procedure. The number of major vessels detected indicate the significance of possible heart problems. “ca” can be categorized into 4 categories respectively: 0-vessel disease, 1-vessel disease, 2-vessel disease and 3-vessel disease.

10. Thalassemia (Thal): Thalassemia refers to the type of blood disorder that is inherited and that affects the hemoglobin production in the patient. Thalassemia is categorized into three categories: Value – 3, Value – 6 and Value – 7. Value 3 thalassemia patients indicate normal and healthy behavior, indicating absence of any heart problems. Value 6 thalassemia patients indicates the presence of fixed defect which means that there is a region in the heart where blood flow is reduced or absent, indicating presence of Heart Disease. Value 7 thalassemia patients indicate that a there exists a region of heart where the blood flow is reduced during stress but reverts back to normal in normal conditions.

3.2 Project Pipeline

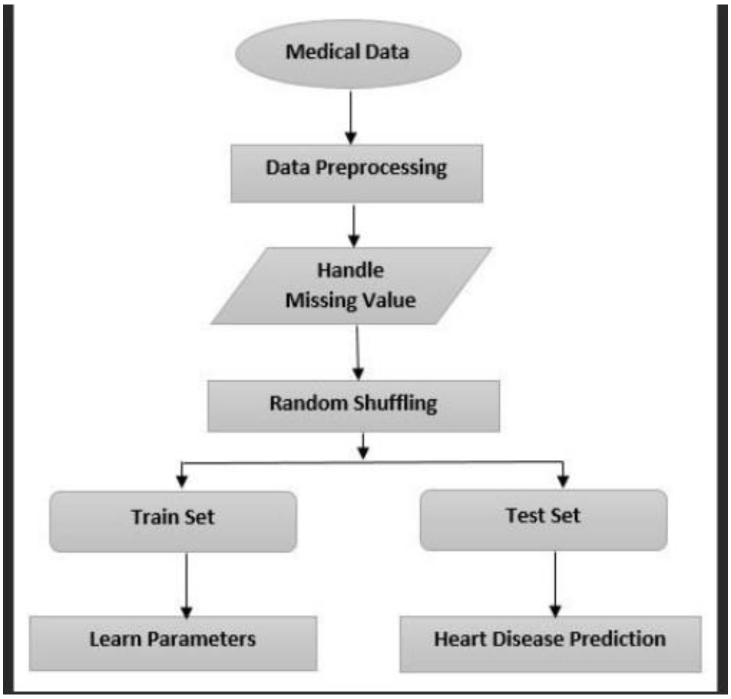
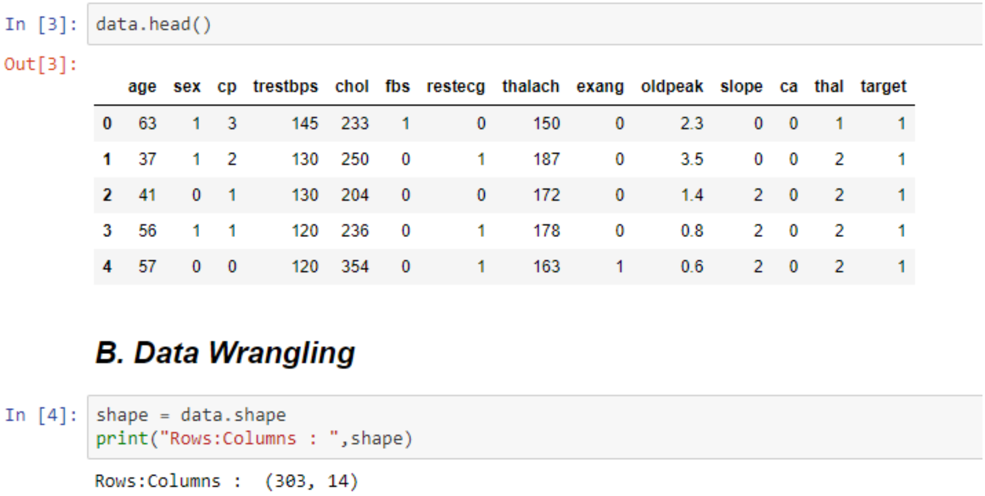


Figure 3.2.1 ML Modeling Pipeline

1. **Data Collection:** After identifying the problem statement for the project or research comes one of the most important steps which completely influences the direction in which the project flows. During the data collection step, we collect data from multiple resources and identify the features that are most crucial and necessary for data analysis and machine learning modeling. For heart disease prediction we need to utilize data that is publicly available in Public Repositories, data that is produced from various clinical trials and studies and the data that is produced from Electronic Health Records. Apart from just collecting data and utilizing it we also need to maintain data ethics by not disclosing the information of the patients whose data we will be utilizing in the project. After considering all the above steps I have compiled a dataset containing 14 crucial features and a total of 303 different values that will help build and train our Machine Learning models.



```
In [3]: data.head()
Out[3]:
```

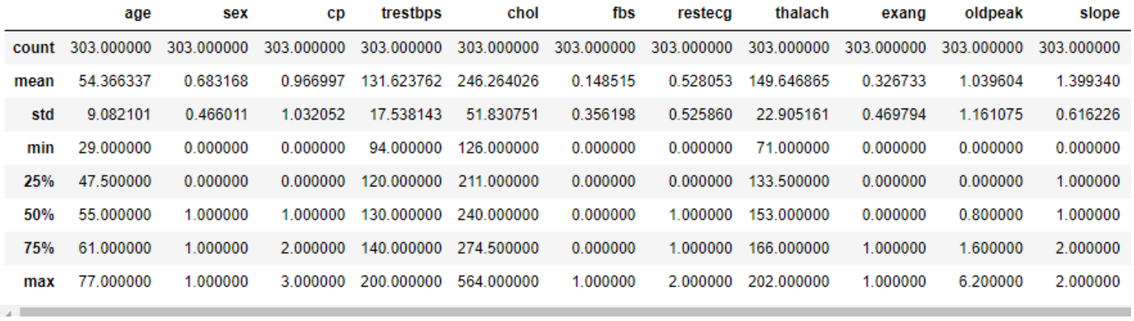
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

B. Data Wrangling

```
In [4]: shape = data.shape
        print("Rows:Columns : ",shape)

Rows:Columns :  (303, 14)
```

Figure 3.2.2 Attributes and Values

2. **Data Preprocessing:** During this stage I extracted crucial information such as size of dataset, number of missing values / null values, number of unique values present in each attribute, statistical measures for each and every attribute and Target value counts for the target column. By doing so we can eliminate any and every possible factor that may influence the decision-making process of the model and prevent any misleading decisions or conclusions made during the process of analysis.
3.



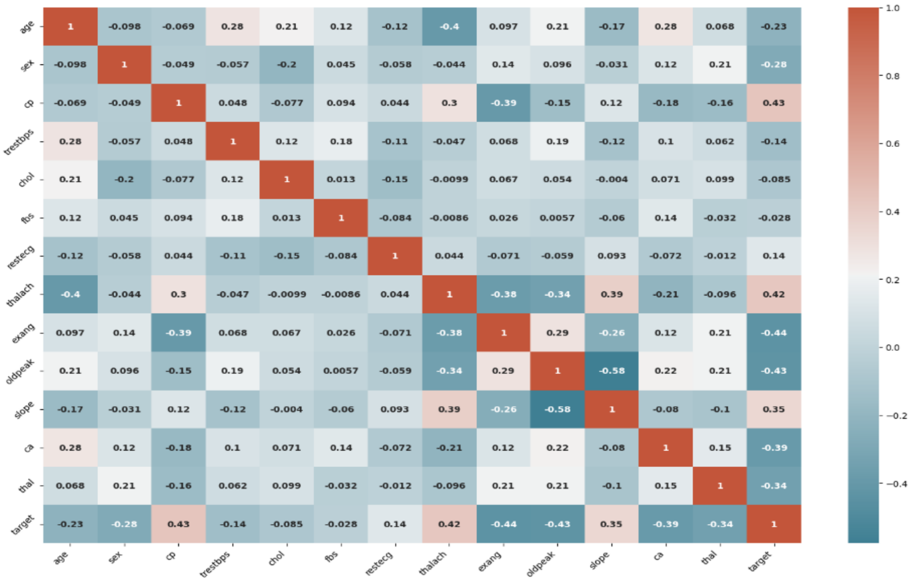
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	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
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	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
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000

Figure 3.2.3 Statistical Measure

4. **Exploratory Data Analysis:** Involves conducting correlation analysis and producing measures such as correlation coefficients that describe the relationships between each attribute and factor. Also includes producing graphs such as Violin Plots, Box Plots, Bar Graphs of various attributes describing significance and contribution of attributes towards the decision making of Machine Learning Models.

    a. ***Analysis of Attributes:*** Conduct the value counts of different values in each attribute, produce a graph related to the value counts and a graph describing the contribution of each value towards the target variable.

    b. ***Correlation Analysis:*** Calculate and create a heat map portraying the correlation coefficient score from values -1 to 1 and describing the relationship between the 13 attributes and the Target attribute. Create Pair plots to observe relationships roughly in the form of visualizations and create violin and box plots to visualize relationship between ST depression , Thalac and Heart Disease.



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1	-0.098	-0.069	0.28	0.21	0.12	-0.12	-0.4	0.097	0.21	-0.17	0.28	0.068	-0.23
sex	-0.098	1	-0.049	-0.057	-0.2	0.045	-0.058	-0.044	0.14	0.096	-0.031	0.12	0.21	-0.28
cp	-0.069	-0.049	1	0.048	-0.077	0.094	0.044	0.3	-0.39	-0.15	0.12	-0.18	-0.16	0.43
trestbps	0.28	-0.057	0.048	1	0.12	0.18	-0.11	-0.047	0.068	0.19	-0.12	0.1	0.062	-0.14
chol	0.21	-0.2	-0.077	0.12	1	0.013	-0.15	-0.0099	0.067	0.054	-0.004	0.071	0.099	-0.085
fbs	0.12	0.045	0.094	0.18	0.013	1	-0.084	-0.0086	0.026	0.0057	-0.06	0.14	-0.032	-0.028
restecg	-0.12	-0.058	0.044	-0.11	-0.15	-0.084	1	0.044	-0.071	-0.059	0.093	-0.072	-0.012	0.14
thalach	-0.4	-0.044	0.3	-0.047	-0.0099	-0.0086	0.044	1	-0.38	-0.34	0.39	-0.21	-0.096	0.42
exang	0.097	0.14	-0.39	0.068	0.067	0.026	-0.071	-0.38	1	0.29	-0.26	0.12	0.21	-0.44
oldpeak	0.21	0.096	-0.15	0.19	0.054	0.0057	-0.059	-0.34	0.29	1	-0.58	0.22	0.21	-0.43
slope	-0.17	-0.031	0.12	-0.12	-0.004	-0.06	0.093	0.39	-0.26	-0.58	1	-0.08	-0.1	0.35
ca	0.28	0.12	-0.18	0.1	0.071	0.14	-0.072	-0.21	0.12	0.22	-0.08	1	0.15	-0.39
thal	0.068	0.21	-0.16	0.062	0.099	-0.032	-0.012	-0.096	0.21	0.21	-0.1	0.15	1	-0.34
target	-0.23	-0.28	0.43	-0.14	-0.085	-0.028	0.14	0.42	-0.44	-0.43	0.35	-0.39	-0.34	1

Figure 3.2.4 Correlation Heatmap

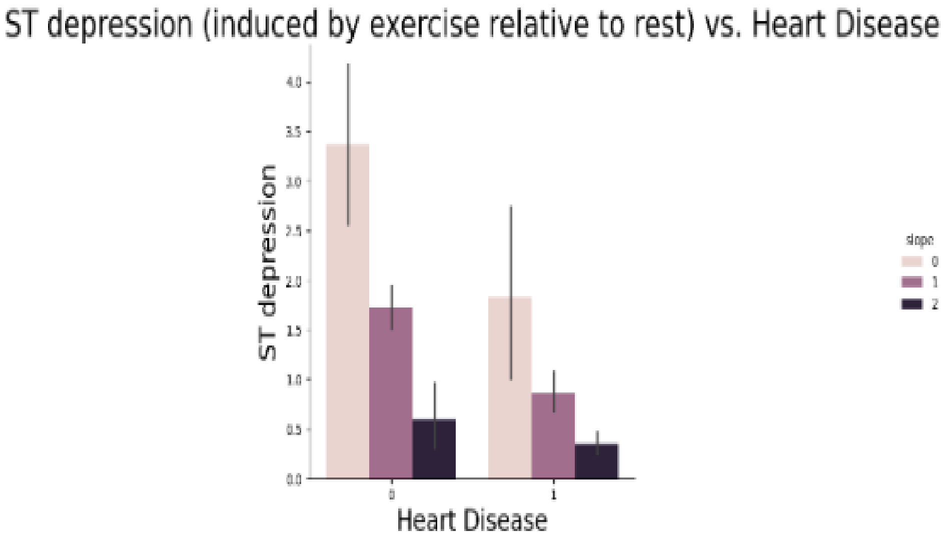


Figure 3.2.5 ST Depression Graph

5. **Machine Learning Algorithms:** Machine learning algorithms are increasingly utilized for predicting heart disease by analyzing vast amounts of medical data and identifying patterns that may not be apparent to human clinicians. By leveraging features such as age, sex, blood pressure, cholesterol levels, and other clinical indicators, machine learning models like logistic regression, decision trees, support vector machines, and neural networks can accurately predict the likelihood of heart disease. These models undergo rigorous training and evaluation to ensure they can generalize well to new, unseen data.

Prepare Data for Model:

Data prepared for Machine Learning models undergoes training and test splitting. The dataset is divided into two subsets namely: training set and the testing set. Typically, we allot 70-80% of the dataset into the training set and the remaining 20-30% of the data to the testing set. This split ensures that the model can generalize well to new data and prevent overfitting of model while making predictions.

Machine Learning Algorithms:

Machine Learning is a vast field of Artificial Intelligence that focuses on developing algorithms that enable computers to learn from and make predictions based on data. It encompasses the use of various techniques to implement algorithms for a range of purposes such as making predictions based on historical or existing data, grouping of similar data based on their attributes, predicting future outcomes by analyzing current patterns and trends, analyze and interpret human language and many such other applications.

Heart Disease Prediction project typically encompasses the use of several Machine Learning algorithms to make predictions.

1. Logistic Regression: Regression analysis models are widely used machine learning algorithms for heart disease prediction due to their simplicity and effectiveness in binary classification tasks. It models the relationship between the set of input features such as age, sex, blood pressure, cholesterol levels etc..., and the probability of a patient having heart disease. In the project using the sklearn library we first import Logistic Regression model as well as the metrics for determining the accuracy of the Logistic Regression model.

2. K-NN: The K-Nearest Neighbors model is a straightforward and intuitive method for heart disease prediction. It classifies the patients health status by comparing their health data such as age, sex, blood pressure, blood sugar levels… to that of other patients. The model identifies the 'K' closest patients (neighbors) based on the similarity of features and predicts possibility of heart disease risk based on the majority class among these neighbors.

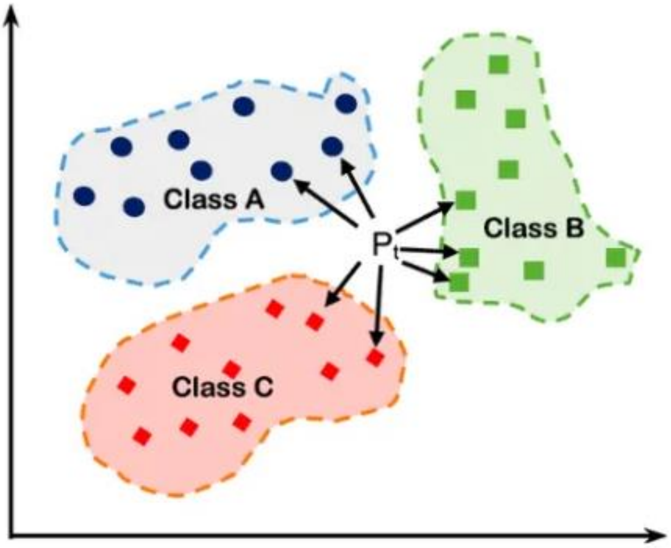


Figure 3.2.6 K-NN distribution graph

3. SVM: The Support Vector Machine model is a powerful tool for heart disease prediction, particularly effective for complex and high-dimensional datasets. Our dataset is a high dimensional dataset since we have included the use of 14 different attributes. SVM works by finding the optimal hyperplane that best separates the patients with and without heart disease based on their features. The hyperplane is a decision boundary that separates different datapoints and we generally aim to maximize the margin (distance between hyperplane and closest datapoints from each class) between the classes so that the model performs consistently and accurately while predicting outcomes.

4. Naïve Bayes Classifier: The Naive Bayes Classifier is a probabilistic model used for heart disease prediction that relies on Bayes' theorem and the assumption of feature

independence. Despite its simplicity, the Naive Bayes Classifier performs well in practice by calculating the probability of heart disease given various patient features like age, blood pressure, and cholesterol levels. It assumes that these features are independent of each other, simplifying the computation of probabilities. This model is particularly efficient for large datasets and can quickly classify patients into those likely to have heart disease and those who are not. Its ease of implementation and speed make it a practical choice for early detection and risk assessment in clinical settings.

5. Decision Trees: The Decision Trees model is a versatile and interpretable method used for heart disease prediction. It works by segmenting the data based on different patient characteristics like age, sex, blood pressure, and cholesterol levels, creating a tree-like structure of decisions. Each node in the tree represents a feature, and each branch represents a decision based on that feature. By recursively partitioning the data, the model can effectively classify patients into groups with different probabilities of having heart disease. Decision Trees are beneficial because they allow healthcare professionals to understand the reasoning behind each prediction, making it easier to explain and trust the model's outcomes.

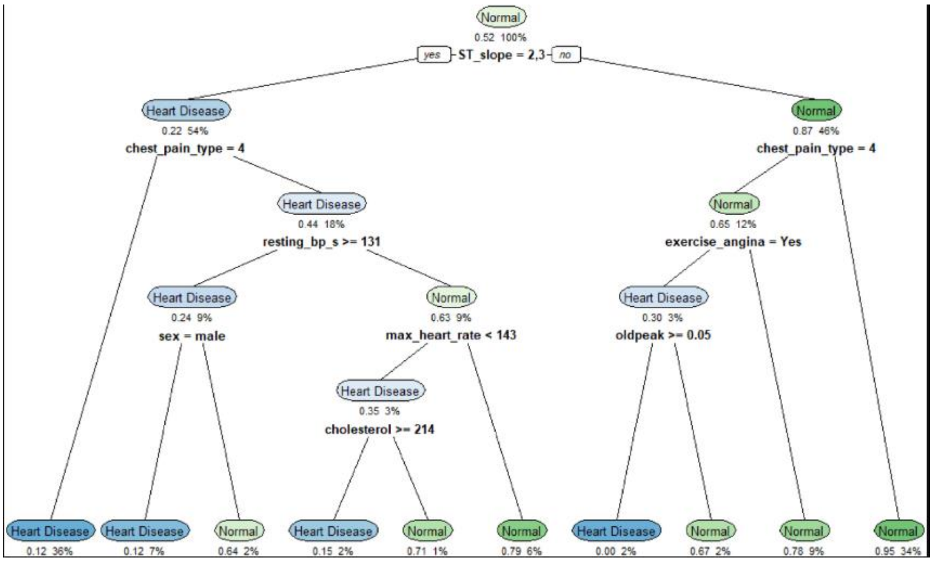


Figure 3.2.7 Decision Trees

6. Random Forest: The Random Forest model is a powerful ensemble learning technique widely used for heart disease prediction. It operates by constructing multiple decision trees during training and outputs the average prediction of those trees. Each tree is trained on a random subset of the data and features, ensuring diversity and reducing overfitting. Random Forests excel in handling high-dimensional datasets with complex interactions between features such as age, sex, blood pressure, and cholesterol levels. They offer robust performance by aggregating predictions from multiple trees, resulting in improved accuracy and generalization of new data. This makes Random Forests a preferred choice for heart disease prediction where reliability and interpretability are crucial.

7. XG Boost: The XG Boost (Extreme Gradient Boosting) model is a machine learning algorithm widely used for heart disease prediction due to its exceptional performance and scalability. It belongs to the ensemble learning family and works by sequentially building a series of decision trees, where each subsequent tree corrects the errors made by the previous one. XG Boost optimizes the model's predictive accuracy by focusing on the instances that are more challenging to classify, improving overall performance. It handles both numerical and categorical data efficiently.

8. Neural Networks: Neural Networks (NN) are advanced machine learning models capable of learning intricate patterns and relationships from complex datasets. They consist of layers of interconnected nodes (neurons) that process input data, learn representations, and produce output predictions. NNs excel in capturing nonlinear relationships among features such as age, sex, blood pressure, and cholesterol levels, which are crucial for accurate diagnosis. Through training on large amounts of medical data, NNs adapt their internal weights to minimize prediction errors, optimizing their ability to generalize to new, unseen patient cases.

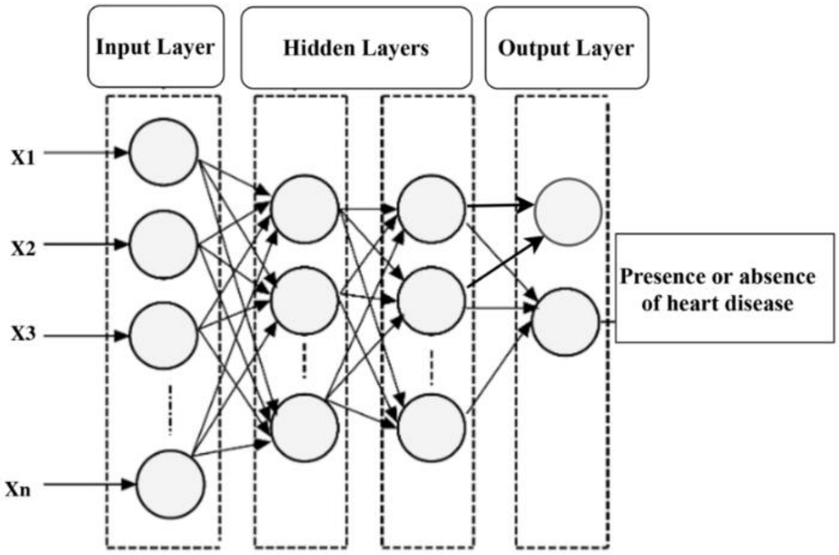
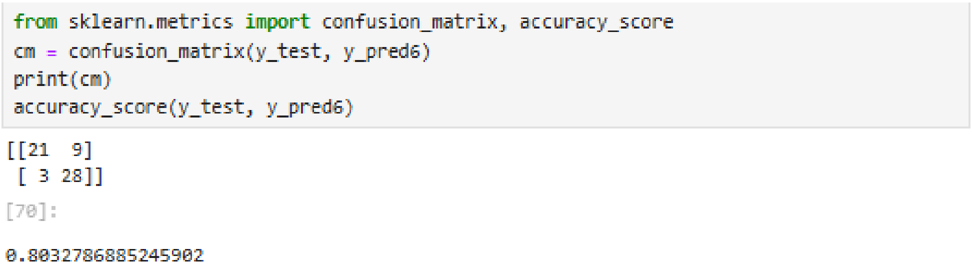


Figure 3.2.8 Neural Network

6. Accuracy Evaluation

Accuracy evaluation is crucial in assessing how effectively a heart disease prediction model performs. It provides a clear measure of the model's ability to correctly classify patients into those with or without heart disease based on their data. This evaluation helps healthcare professionals understand the model's reliability and performance in real-world scenarios, guiding decisions on its implementation and use in clinical practice.

For conducting accuracy evaluation, I will be utilizing Confusion Matrix which is a critical tool for providing a summary of the models performance by comparing predicted outcomes against actual outcomes. The matrix is structured into four quadrants: true positives (correctly predicted positive outcomes), true negatives (correctly predicted negative outcomes), false positives (incorrectly predicted positive outcomes), and false negatives (incorrectly predicted negative outcomes). By analyzing these metrics, such as sensitivity (true positive rate), specificity (true negative rate), precision, and recall, healthcare professionals can assess how well the model identifies patients with heart disease and those without. This evaluation helps in understanding the model's strengths and weaknesses, guiding further improvements or adjustments to enhance its diagnostic accuracy.



```
from sklearn.metrics import confusion_matrix, accuracy_score
cm = confusion_matrix(y_test, y_pred6)
print(cm)
accuracy_score(y_test, y_pred6)
```

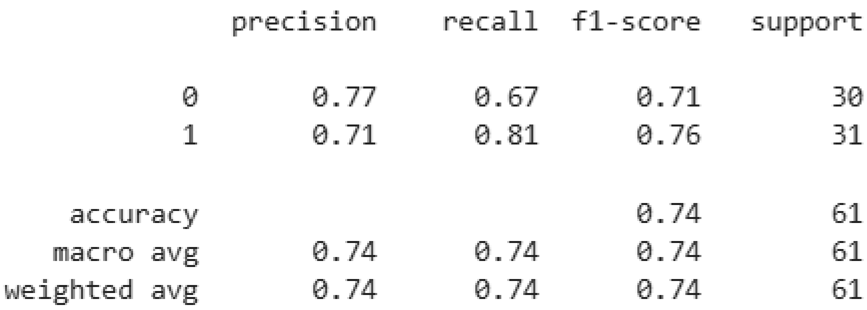
```
[[21  9]
 [ 3 28]]
[70]:

0.8032786885245902
```

Figure 3.2.9 Accuracy Evaluation

4. Implementation

4.1 Machine Learning Modeling

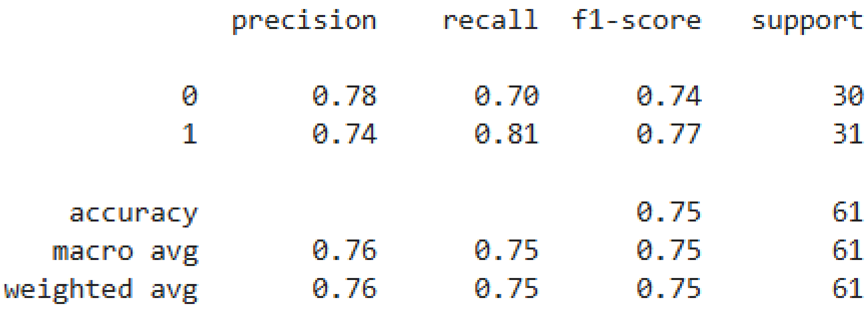
1. Logistic Regression: We then fit the training data to the model and train the model with the same data. Finally, we make predictions from the trained model using the testing dataset and obtain measures such as Precision, Recall, F1 score and support to determine the accuracy of the Machine Learning model. Based on the metrics we observe that the training and test dataset produces a result of **74%** accuracy, which is a favorable result.



```
              precision    recall  f1-score   support

           0       0.77      0.67      0.71        30
           1       0.71      0.81      0.76        31

    accuracy                           0.74        61
   macro avg       0.74      0.74      0.74        61
weighted avg       0.74      0.74      0.74        61
```

Figure 4.1.1 Logistic Regression

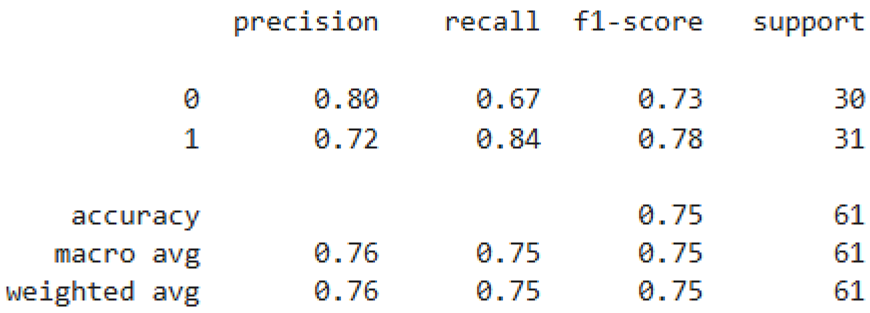
2. K Nearest Neighbors: After training and testing the model using the train and test dataset, we come to the conclusions that the model provides an accuracy of **75%** which is a more favorable outcome compared to Logistic Regression model.



```
              precision    recall  f1-score   support

           0       0.78      0.70      0.74        30
           1       0.74      0.81      0.77        31

    accuracy                           0.75        61
   macro avg       0.76      0.75      0.75        61
weighted avg       0.76      0.75      0.75        61
```

Figure 4.1.2 K-Nearest Neighbors

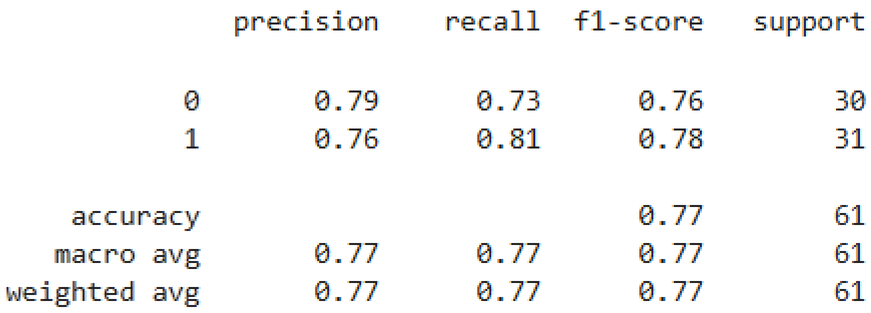
3. Support Vector Machines: After training and testing the model, we achieve an accuracy of **75%,** that means it is equally favorable to that of KNN model.



```
              precision    recall  f1-score   support

           0       0.80      0.67      0.73        30
           1       0.72      0.84      0.78        31

    accuracy                           0.75        61
   macro avg       0.76      0.75      0.75        61
weighted avg       0.76      0.75      0.75        61
```

Figure 4.1.3 K-Nearest Neighbors

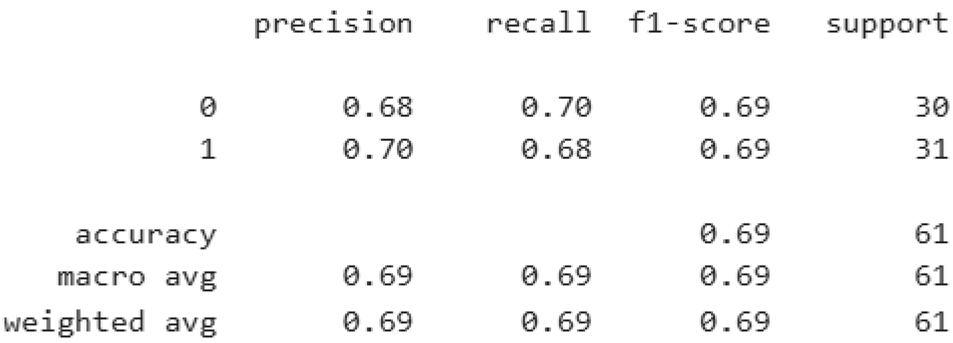
4. Naïve Bayes Classifier: The model after training and testing stages, produces an accuracy of **77%,** which outshines the accuracy levels of other machine learning models, thus making it more favorable for making predictions.



```
              precision    recall  f1-score   support

           0       0.79      0.73      0.76        30
           1       0.76      0.81      0.78        31

    accuracy                           0.77        61
   macro avg       0.77      0.77      0.77        61
weighted avg       0.77      0.77      0.77        61
```

Figure 4.1.4 Naïve Bayes Classifier

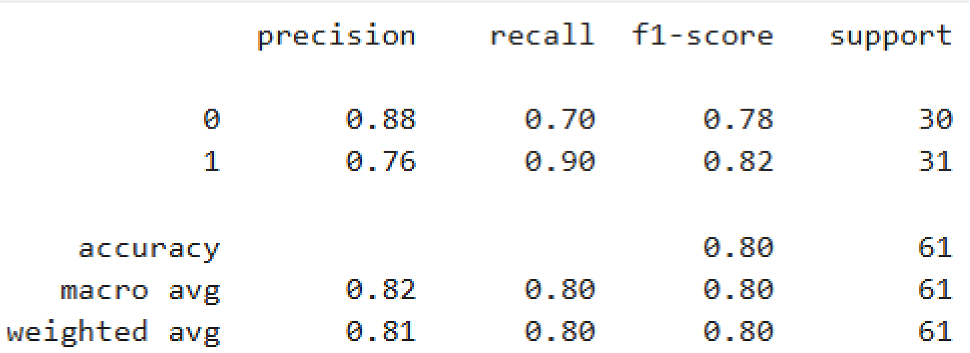
5. Decision Trees: The model produces an accuracy of **69%** while making predictions which in comparison to other models is very less, thus making it less favorable to be used for making predictions.



```
              precision    recall  f1-score   support

           0       0.68      0.70      0.69        30
           1       0.70      0.68      0.69        31

    accuracy                           0.69        61
   macro avg       0.69      0.69      0.69        61
weighted avg       0.69      0.69      0.69        61
```

Figure 4.1.5 Decision Trees

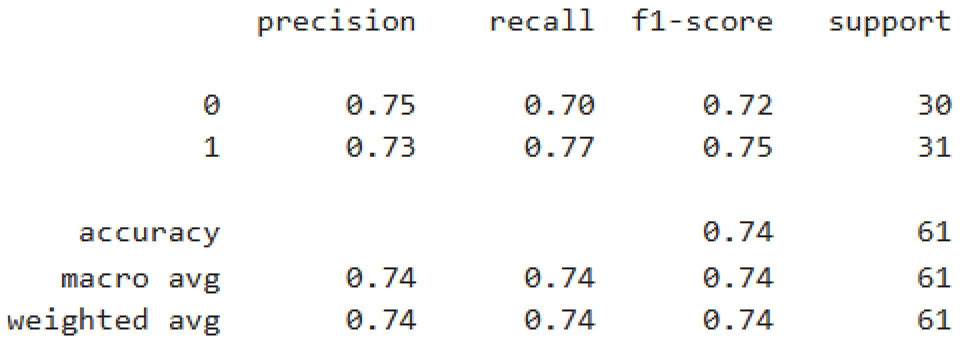
6. Random Forest: The model after training and testing stages, produces an accuracy of **80%,** which marginally outshines the accuracy levels of other machine learning models, thus making it more favorable for making predictions.



```
              precision    recall  f1-score   support

           0       0.88      0.70      0.78        30
           1       0.76      0.90      0.82        31

    accuracy                           0.80        61
   macro avg       0.82      0.80      0.80        61
weighted avg       0.81      0.80      0.80        61
```

Figure 4.1..6 Random Forest

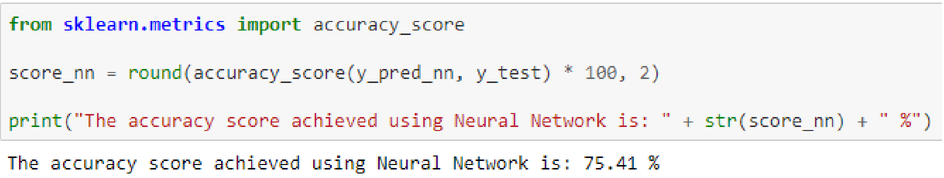
7. XG Boosting: The model provides an accuracy of **74%**, which is less favorable in comparison to that of Random Forest.



```
              precision    recall  f1-score   support

           0       0.75      0.70      0.72        30
           1       0.73      0.77      0.75        31

    accuracy                           0.74        61
   macro avg       0.74      0.74      0.74        61
weighted avg       0.74      0.74      0.74        61
```

Figure 4.1.7 XG Boost

8. Neural Networks: The model after training and testing stages, produces an accuracy of **75%,** which means that the neural networks underperformed as compared to the performance levels of other machine learning models, thus making it less favorable for making predictions.



```
from sklearn.metrics import accuracy_score

score_nn = round(accuracy_score(y_pred_nn, y_test) * 100, 2)

print("The accuracy score achieved using Neural Network is: " + str(score_nn) + " %")
```

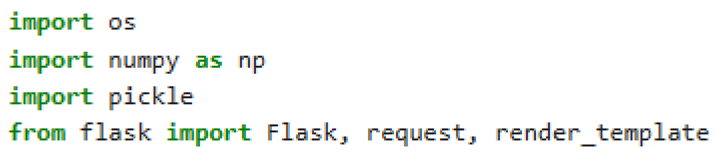
```
The accuracy score achieved using Neural Network is: 75.41 %
```

Figure 4.1.8 Neural Networks

4.2 Flask API

The Flask Application serves the heart disease prediction model via a web interface. Users input their health parameters into a form and upon submission the application uses the machine learning model to predict the likelihood of Heart Disease. The result is then displayed on the web page. The code effectively demonstrates how the Machine Learning model is integrated into the web interface using Flask, making predictions and updates the user interface based on the output of the Machine Learning model. The API is divided into 6 components:

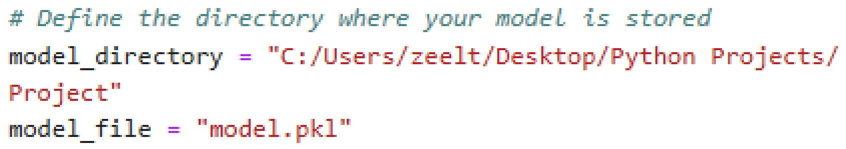
1. **Import Libraries:** Python libraries such as os, NumPy, pickle, flask are imported to perform various functions such as performing numerical operations, existing as a lightweight web framework for Python, provide an interface to interact with the operating system and many more such features.



```
import os
import numpy as np
import pickle
from flask import Flask, request, render_template
```

Figure 4.2.1 Libraries

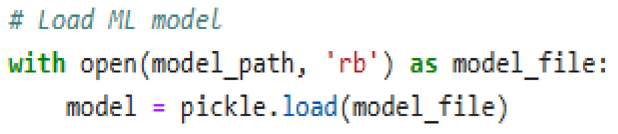
2. **Define Model Directory and Path:** This component is particularly used for specifying the location of the trained machine learning model on the system which helps the web page navigate when any user interacts with the web page.



```
# Define the directory where your model is stored
model_directory = "C:/Users/zeelt/Desktop/Python Projects/
Project"
model_file = "model.pkl"
```

Figure 4.2.2 Model Directory

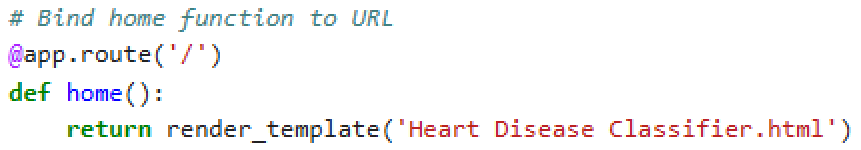
3. **Load Machine Learning Model:** The file containing the model is accessed and the model is opened in read-binary mode and loaded into the variable “model” using “pickle”. By opening the model in read-binary mode and saving it into model variable we deserialize the machine learning model that was previously saved to the disk. This allows the saved model to be reused multiple times without having to train it again whenever we access the model.



```
# Load ML model
with open(model_path, 'rb') as model_file:
    model = pickle.load(model_file)
```

Figure 4.2.3 Load Model

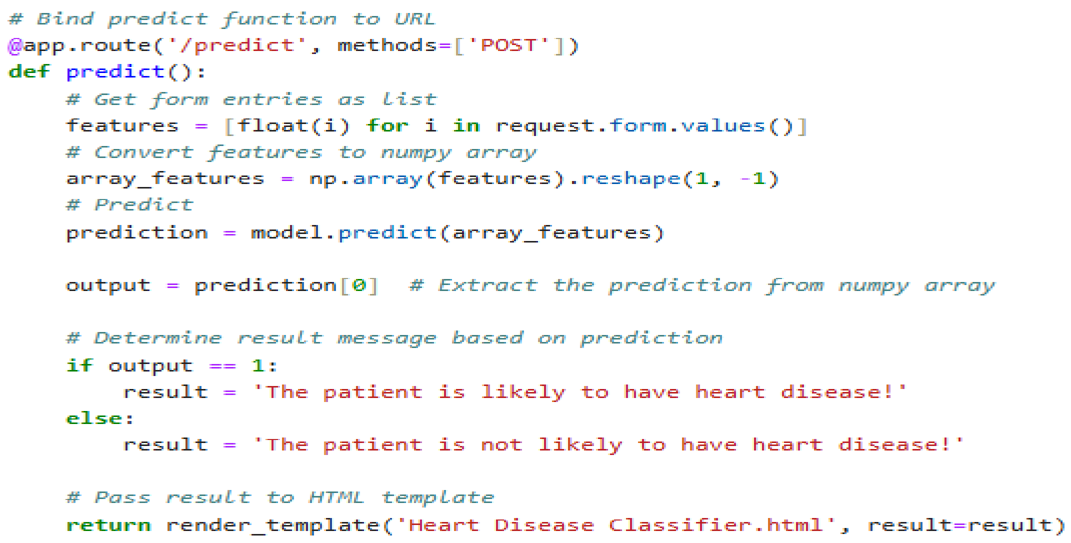
4. **Define Routes:** Routes in flask application are used to map URLs to functions. Each route is associated with a particular URL, and it triggers the function when the URL is being accessed. This allows us to access models when we access the URL for our web interface. The Home route renders the web template which contains the form that will take user input.



```
# Bind home function to URL
@app.route('/')
def home():
    return render_template('Heart Disease Classifier.html')
```

Figure 4.2.4 Binding Home Page

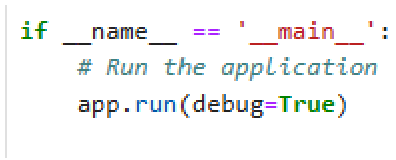
The predict route retrieves the input values from the form submitted by the user and converts them into a list of floats. These features are converted into suitable values that the model can interpret and make a prediction. Based on the prediction produced the results are passed back to the html template.



```
# Bind predict function to URL
@app.route('/predict', methods=['POST'])
def predict():
    # Get form entries as list
    features = [float(i) for i in request.form.values()]
    # Convert features to numpy array
    array_features = np.array(features).reshape(1, -1)
    # Predict
    prediction = model.predict(array_features)

    output = prediction[0]  # Extract the prediction from numpy array

    # Determine result message based on prediction
    if output == 1:
        result = 'The patient is likely to have heart disease!'
    else:
        result = 'The patient is not likely to have heart disease!'

    # Pass result to HTML template
    return render_template('Heart Disease Classifier.html', result=result)
```

Figure 4.2.4.1 Binding Predict function

5. **Run the Application:** The final step of the complete process is to run the Flask Application in debug mode, which helps provide detailed description of error messages and automatically restarts the server whenever any changes to code are made.



```
if __name__ == '__main__':
    # Run the application
    app.run(debug=True)
```

Figure 4.2.5 Run Application

4.3 Web Page

The Web Page is a crucial part of the Heart Disease Prediction project for many reasons. It provides a user-friendly interface for the users to provide their health data and in return derive conclusions from the same using Machine learning models. The web page also facilitates the interaction between the machine learning model and prediction results I provides immediately after the user submits his/her data. A web-based application is also highly accessible since it can be accessed from various devices such as desktops, laptops, tablets and smartphones. The web page is divided into 3 major components:

1. **Head Section:** This section includes links to Bootstrap CSS and JavaScript libraries that are used for styling the web page and adding interactive elements so that the web page is more visually appealing and interactive for the end user.

2. **Body Section:** The body section includes the core content of the code and is the skeletal frame of the web page.

    a. Form: The form collects user inputs such as age, sex, chest pain type, blood pressure, cholesterol levels and other relevant health parameters for the Heart Disease Prediction Model.

    b. Form Container: This container holds the form where users input their health parameters and it is structured using CSS and HTML.

    c. Submit Button: Allows the users to submit the form after all the health parameter fields have been filled. The input data from this section is sent to the Flask backend for further processing and prediction analysis.

    d. Prediction Result Display: Below the Submit button the, web page displays the result of prediction analysis based on the User input, informing them whether they are likely to have a heart disease or not.

3. **Ending Section:** This section is an essential part of the framework of the web page since it ensures that the HTML code is correctly interpreted by the browser and after the web page is closed the backend processes are terminated.

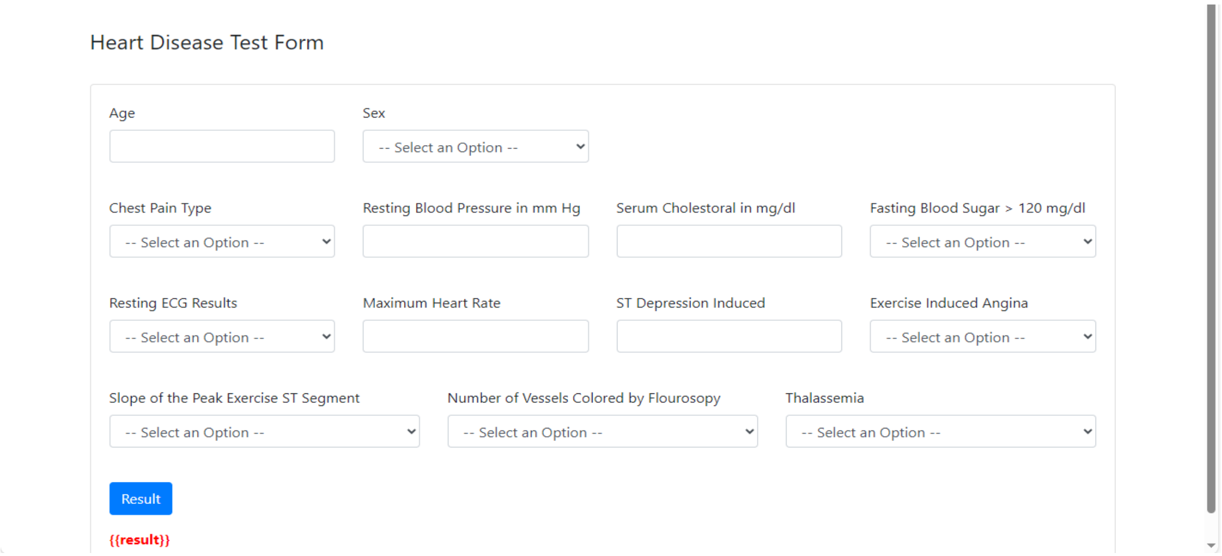


Figure 4.3.1 Web Page Form

5. Results and Analysis

This section comprises of the results of my heart disease prediction models, detailing the performance metrics and key findings from analysis of dataset.

5.1 Exploratory Data Analysis

During the data exploration and analytical stages, several inferences were made regarding the distribution of values in the data set.

a. Sex: Filtering based on target variable, we could deduce conclusions that females in general we potentially at greater risk of contracting heart disease.

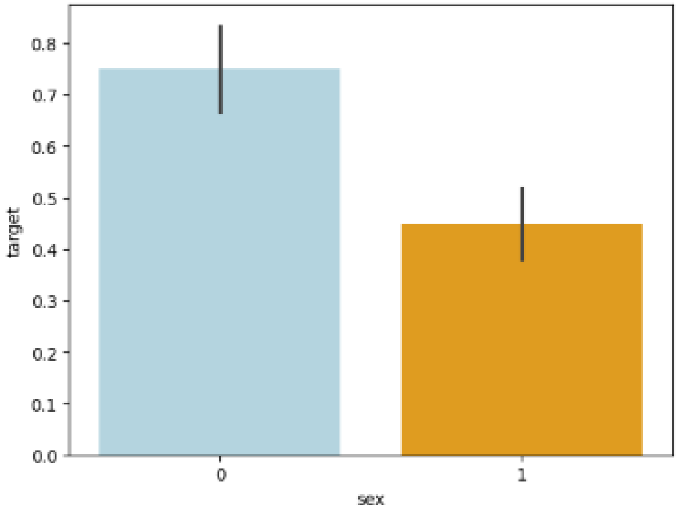


Figure 5.1.1 Sex Distribution

b. Chest Pain: Based on category of chest pain the we could deduce the possibility of a person being a potential Heart Disease patient.

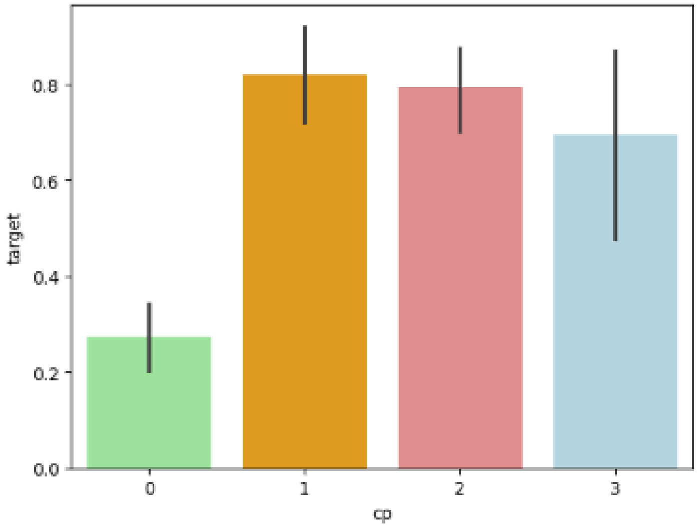


Figure 5.1.2 Chest Pain Distribution

c. Rest ECG: People displaying Rest ECG values as 0 and 1 are more likely to develop heart disease compared to those having value 2.

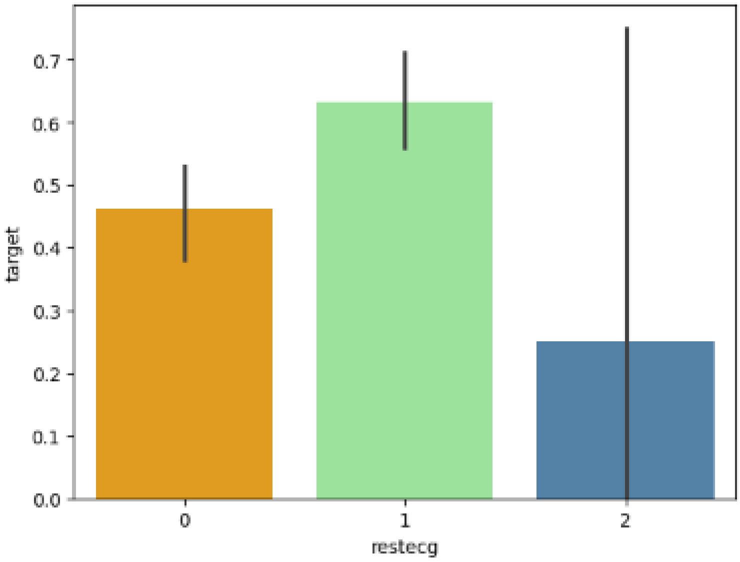


Figure 5.1.3 Rest ECG

d. Coronary Artery (ca): Patients having ca value 4 are the people who are having the highest possibility of having Heart Disease, followed by are the patients who are having ca value 0. Patients with ca value 1,2 and 3 are the least susceptible patients.

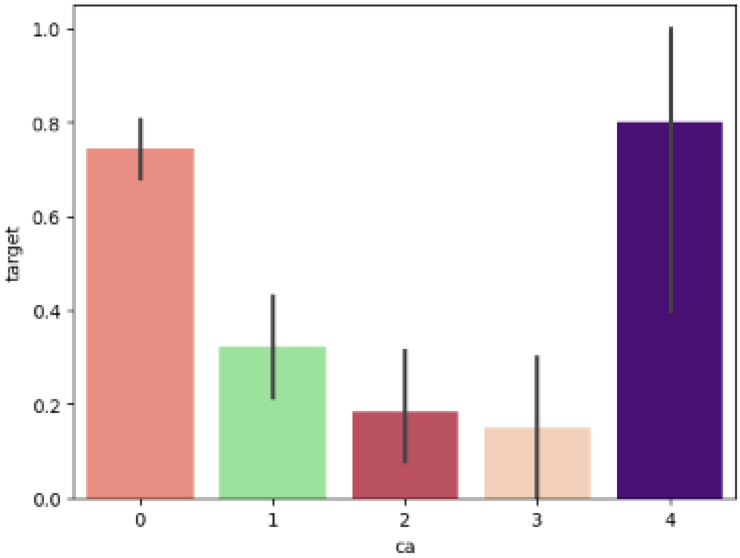


Figure 5.1.4 Coronary Artery

e. Thalassemia (Thal): Patients having thalassemia value 2 are more likely to contract or have heart disease compared to those having values 0, 1 and 3.

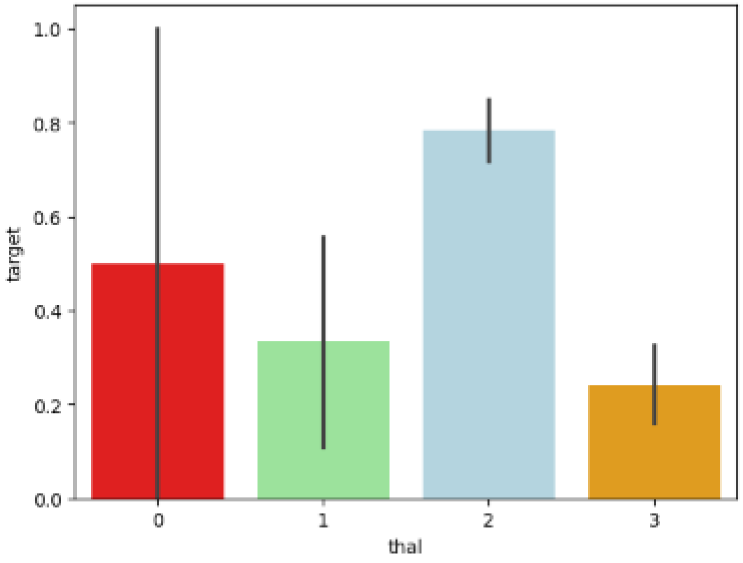


Figure 5.1.5 Thalassemia

After a thorough process of performing data exploration and initial analysis we came to the conclusion using Correlation analysis that there are 4 important features that are directly related to the possibility of a patient being a Heart Disease target. By visualizing the correlation analysis in the form of correlation matrix we calculate the contribution of each feature towards the patient being the Target variable for Heart Disease.

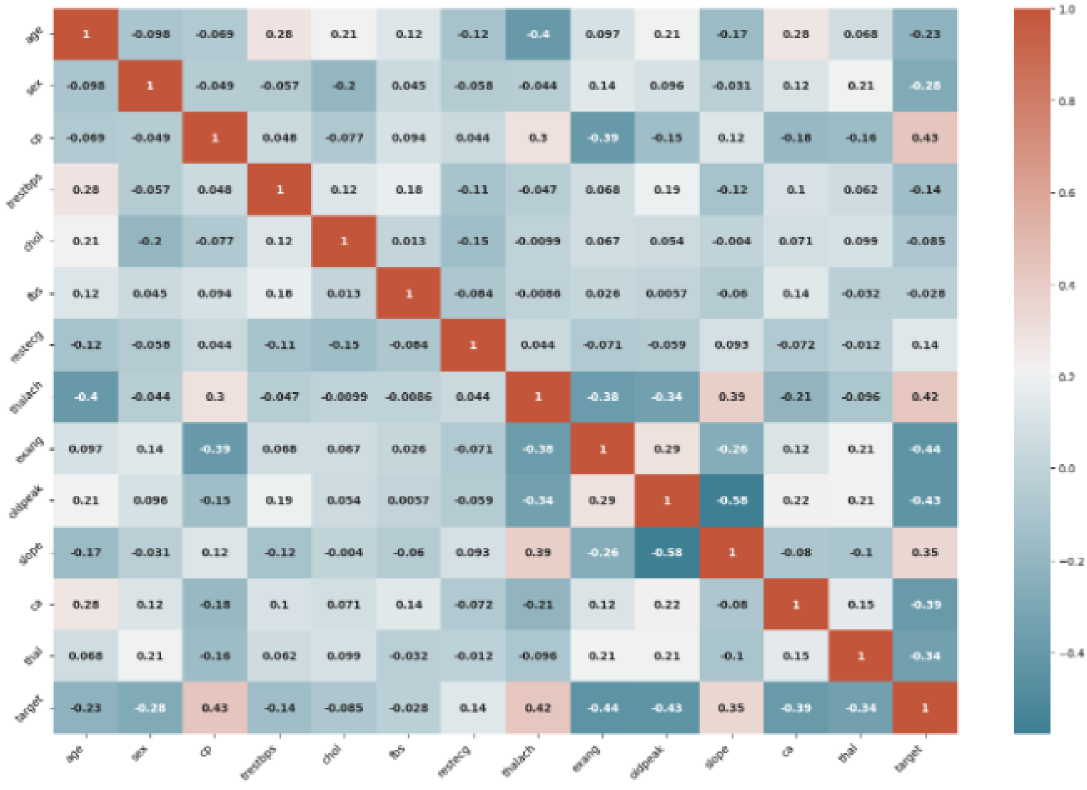
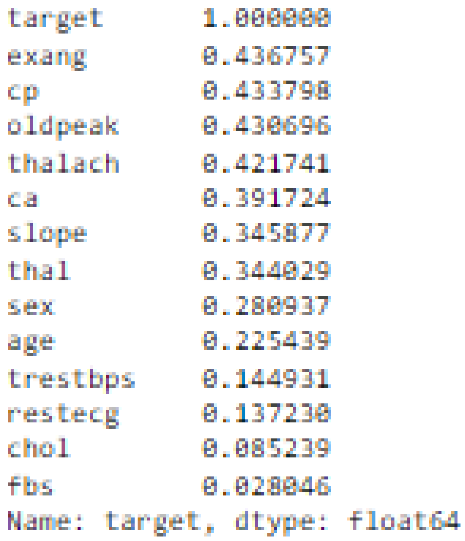


Figure 5.1.6 Correlation Matrix



```
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
```

Figure 5.1.7 Correlation Values

5.2 Machine Learning Model

To produce accurate and effective results, the project utilizes the use of multiple machine learning models. A total of seven different Machine Learning models we used and one deep learning model. Out of all the models Random Forest produced results with the most accuracy after training and testing it using the Train-Test split data. It produced an accuracy of eighty percent which is considered as a good accuracy for a machine learning model. It shows that there is neither overfitting nor underfitting. Rest other models produced similar results of accuracy ranging between seventy to seventy-five percent.

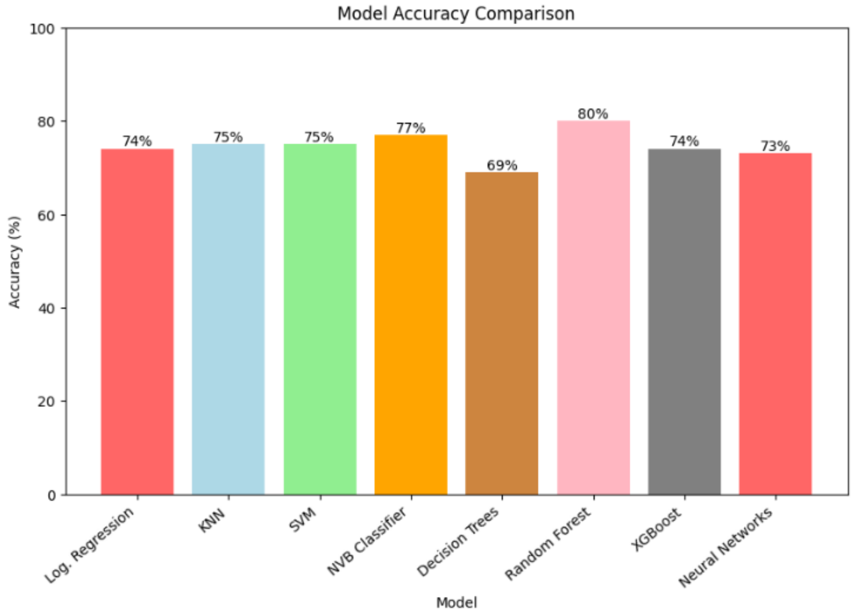
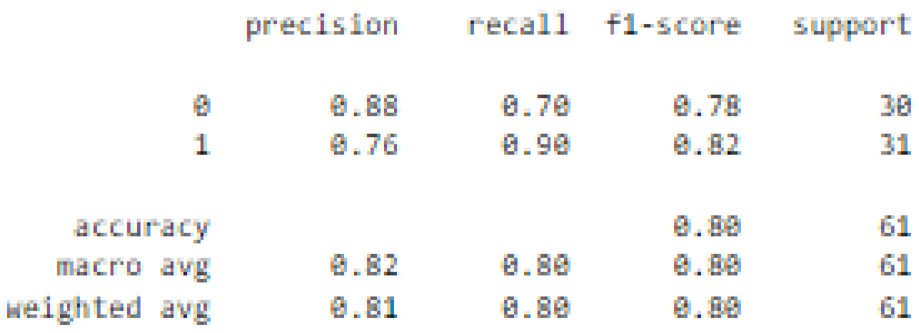


Figure 5.2.1 Accuracy Comparison

Apart from overall high value of accuracy the Random Forest model performed well in other forms of metrics as well. These metrics include Precision value, F1 Score, Recall values and Support values. It produced values that are considered to be one of the highest amongst all the models and they are 0.76, 0.82 and 0.90 respectively.



```
              precision    recall  f1-score   support

           0       0.88      0.70      0.78        30
           1       0.76      0.90      0.82        31

    accuracy                           0.80        61
   macro avg       0.82      0.80      0.80        61
weighted avg       0.81      0.80      0.80        61
```

Figure 5.2.2 Accuracy Metrics

The performance of the Random Forest model can also be visualized in the form of ROC curve and Precision Recall curve. The ROC curve plots the True Positive Rate against the False Positive Rate at various thresholds. The Area Under the Curve (AUC) is used to determine the models ability to distinguish between the two rate classes. If AUC of the curve is calculated as 1 then the ML model indicates perfect performance. If the AUC of the curve is calculated as 0.5 or less than 0.5 then the model indicates indistinguishable performance. Our Random Forest model achieved a score of 0.74 in terms of AUC which is greater than 0.5, hence we can conclude that the model indicates high levels of performance.

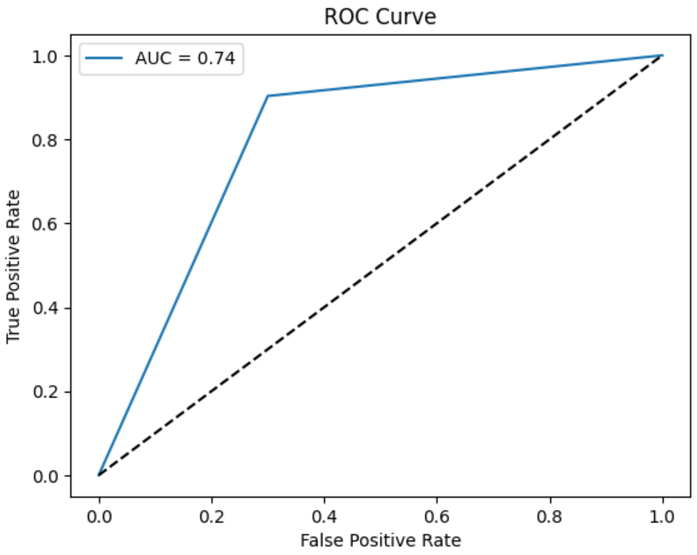


Figure 5.2.3 ROC curve

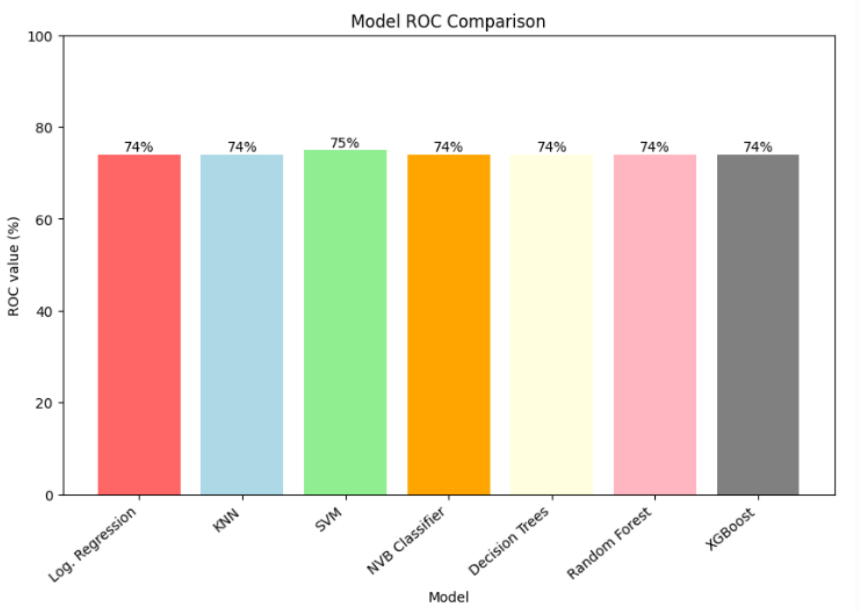


Figure 5.2.4 ROC comparison

The precision-recall graph plots the precision (positively predicted values) against recall (sensitivity) at various thresholds. The Precision Recall curve is useful for assessing the models model's ability to identify positive cases without generating too many false positives. A high area under the curve (AUC) indicates both high precision and high recall, which generally means the model is performing well. In the case of the Random Forest model, the value achieved for AUC is 0.85 or 85% which in comparison to other Machine Learning models is of highest value. Thus, our model is more accurate and well balanced for making predictions.

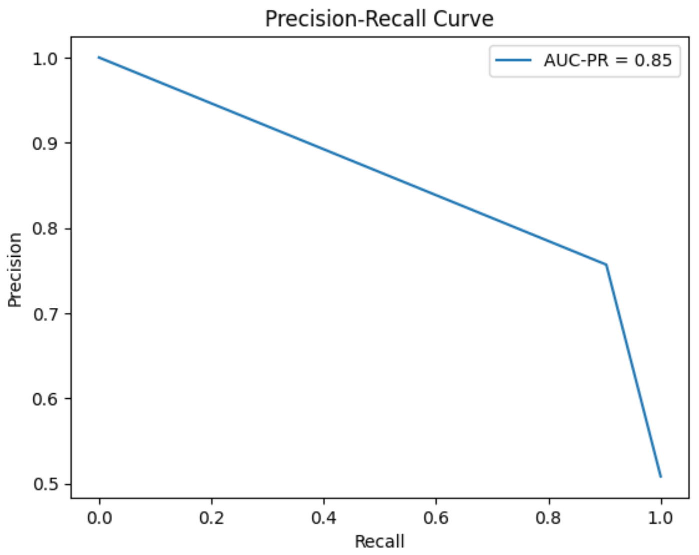


Figure 5.2.5 Precision Recall Curve

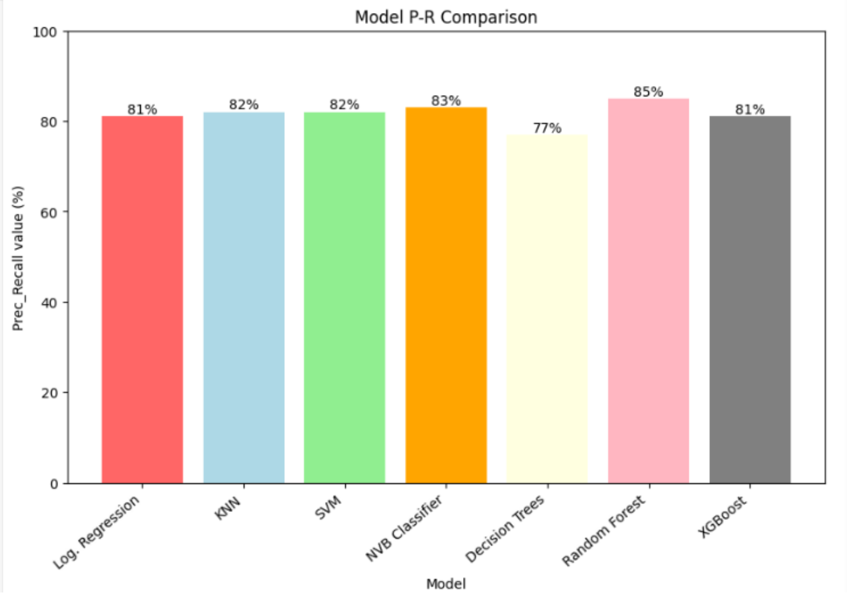
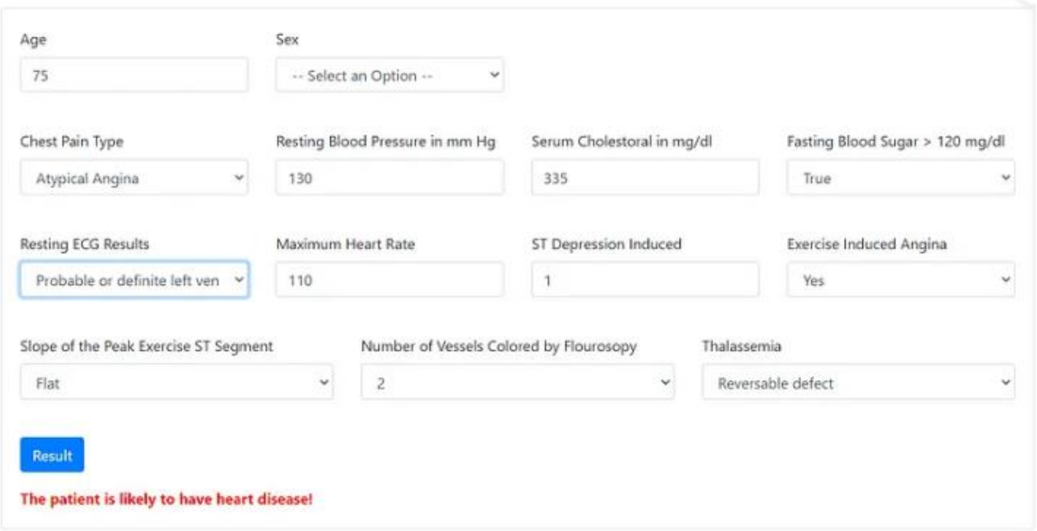


Figure 5.2.6 P-R Comparision

5.3 Heart Disease Classifier

The heart disease classifier is the web page of our Heart Disease Prediction project. It is web page that takes user input information such as Age, Sex, Serum Cholesterol, ST depression value, and many other such features as input for the Machine learning model. The random forest model that provides us with an accuracy of 80% is attached to the web page using Flask API and runs in the background for analysis and prediction. Upon entering values, the model produces results after analysis and the result is reflected back on the Web Page below the Submit button. Two types of output can be produced depending on the type of values entered by the user: 1. The patient is not likely to have heart disease, 2. The patient is likely to have heart disease.



Age
75

Sex
-- Select an Option --

Chest Pain Type
Atypical Angina

Resting Blood Pressure in mm Hg
130

Serum Cholestoral in mg/dl
335

Fasting Blood Sugar > 120 mg/dl
True

Resting ECG Results
Probable or definite left ven

Maximum Heart Rate
110

ST Depression Induced
1

Exercise Induced Angina
Yes

Slope of the Peak Exercise ST Segment
Flat

Number of Vessels Colored by Flourosopy
2

Thalassemia
Reversable defect

The patient is likely to have heart disease!

Figure 5.3.1 Web Page Result

6. *Conclusion and Future Plan*

Machine Learning has revolutionized the Healthcare industry by leveraging data driven insights to enhance treatment effectiveness and diagnostic accuracy. In the realm of Heart Diseases, ML algorithms effectively analyze vast datasets comprising of multiple number of features like patient demographics, clinical parameters and lifestyle factors to identify underlying patterns and predict the likelihood of a patient having Heart Disease.

Our project intends to prove the same, the Random forest model amongst various other models stands out the most. It yields an accuracy of 80% where any accuracy above 70% is considered to be good. Out of the 13 features we examined the top 4 significant features that helped us classify between positive and negative diagnosis were Chest Pain type (cp), Maximum Heart Rate achieved (Thalach), Number of Major vessels (ca) and ST depression induced by exercise relative to rest (oldpeak). The below diagram shows the same;

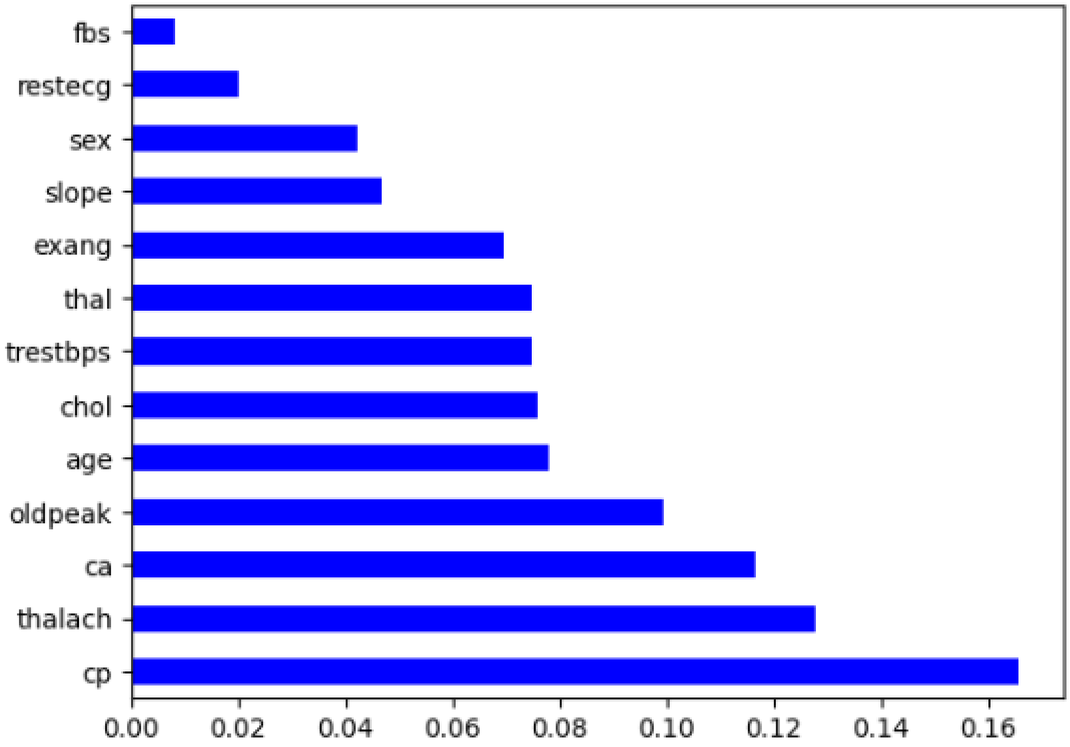


Figure 6.1 Feature Importance

Our Machine Learning algorithm can now classify patients with Heart Disease, and we can properly diagnose patients and get them the help and treatment they need to recover. By early diagnosis we can prevent any worse symptoms from arising later.

The future of Heart Disease Prediction holds immense potential. As Machine Learning models continue to evolve and upgrade, we expect them to refine their predictive capabilities by integrating more diverse data sources and leverage learning capabilities from algorithms like Deep Learning. This not only will improve accuracy but also help provide personalized risk assessments that are tailored to individual patient profiles. For future purposes we can also tailor treatments and preventive measures based on individual risk profiles and severity derive by advanced predictive models.

Heart Disease Prediction can also be implemented into real time analysis and similar concepts. We can leverage data produced from wearable devices and IoT sensors that continuously monitor heart health metrics. Machine Learning algorithms can analyze this data in real time to provide early signs of warnings and recommendations to the end user.

Overall, the future of Heart Disease lies in harnessing the power of Advanced Techniques such as Deep Learning and Artificial Intelligence and interdisciplinary collaboration to enhance early detection, personalized treatment suggestions and ultimately improve the health of patients that are or maybe at risk of cardiovascular diseases.

7. *References*

1. ***UCI Dataset: Heart Disease - UCI Machine Learning Repository***

2. *Noncommunicable Diseases Country Profiles. World Health Organization; 2018. https://www.who.int/nmh/publications/ncd-profiles-2018/en/ [Internet] 2019 [cited 17 December 2019]. Available from: [Google Scholar]*

3. *Institute for Health Metrics and Evaluation (IHME). Findings from the Global Burden of Disease Study 2017. IHME; Seattle, WA: 2018. http://www.healthdata.org/sites/default/files/files/policy_report/2019/GBD_2017_Booklet.pdf;2019 [Internet]. Healthdata.org [cited 17 December 2019] Available from: [Google Scholar]*

4. *Kasthuri A. Challenges to healthcare in India - the five A's. Indian J Community Med. 2018;43(3):141–143. doi: 10.4103/ijcm.IJCM_194_18. [PMC free article] [PubMed] [CrossRef] [Google Scholar]*

5. *Ramalingam VV, Dandapath A, Raja MK. Heart disease prediction using machine learning techniques: a survey. Int J Eng Technol. 2018;7(2.8):684–7. doi: 10.14419/ijet.v7i2.8.10593.*

6. *Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE. 2017;12(4) doi: 10.1371/journal.pone.0174944.*

7. *Anna Karen Garate-Escamila, Amir Hajjam El Hassani, Emmanuel Andres, Classification models for heart disease prediction using feature selection and PCA, Elsevier Informatics in Medicine Unlocked, April 2020.*