Heart Care Heart Attack Possibility

Shahajadi Sadia Afsana 180104138

Project Report
Course ID: CSE 4214
Course Name: Pattern Recognition Lab
Semester: Spring 2021

Department of Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh

August 2022

Heart Care Heart Attack Possibility

Submitted by

Shahajadi Sadia Afsana **180104138**

Submitted To

Faisal Muhammad Shah

Sajib Kumar Saha Joy,

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

August 2022

ABSTRACT

The heart seems to be a very complicated organ in the human body. If some part of the heart has been seriously damaged, the remaining part of the heart will remain functioning. But as a result of the injury, the heart can be weakened and unable to pump as much blood as normal. With timely detection of multiple possible hamstring issues, proper care, and dietary changes after a heart attack, the additional injury can be reduced or avoided. In this paper, different types of machine learning algorithms are used for measuring the possibility of heart attack, they are random forest, MLP, KNN and Adaboost. By finding the best algorithm, this paper also shows the confusion matrices, visualizes the feature, and ROC curve. From this research work, it is evident that KNN the best model with an accuracy of about 82% and also gives the best Roc of about 82%.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cardiovascular disorder refers to a group of heart-related diseases. Cardiovascular disease encompasses blood vessel disorders such as coronary artery disease, heart rhythm disorders, and heart defects such as atrial fibrillation (congenital heart defects). The heart is our body's most vital organ. The heart is the lifeline of our bodies. Heart disease is a common condition. A heart attack is one of the most common illnesses. When a heart attack occurs, the blood supply does not function properly. Following a heart attack, fat, cholesterol, and other compounds are commonly accumulated. An attack, or myocardial infarction, occurs when a portion of the cardiac muscle does not receive enough oxygen. Symptoms of heart disease include abdominal pain or pressure. Many heart attacks cause emotional or left chest pain that lasts for more than a few minutes or leaves and returns. Uncomfortable strains, rubbing, fullness, or pain can all be unpleasant. Weakness, dizziness, or faintness You might also break out in a cold sweat. Irritation or pressure on the mask, neck, or back. Pain or irritation in one or both arms or shoulders. Shortening of the breath This can cause heart pain, but it can also cause a lack of oxygen prior to a heart attack. Heart disease statistics: Last year, 735,000 Americans suffered a heart attack. Cardiovascular Disease is the leading cause of death in the United States. Only about 0.3% of men and women aged 20 to 39 have heart disease. The average age of the first heart attack is 65 years for men and 72 years for women.

Factor of Risk: The following are the most important risk factors for heart attack control: Alter. Alter. Men over 45 and women over 55 are more likely to have a heart attack than younger men and women. Smoking. Smoking. Smoking and second-hand smoke exposure were both considered. Blood pressure is elevated. High blood pressure can cause artery damage to your heart over time. Obesity, high cholesterol, or diabetes all increase the chances of developing high blood pressure. Serum or triglyceride cholesterol levels are higher. Low density lipoprotein may be linked to high cholesterol ("bad" cholesterol) (LDL). A high level of triglycerides, a type of blood fat associated with diet, increases your

risk of having a heart attack. A high volume of high-density cholesterol ("good" cholesterol") lipoprotein would reduce the risk (HDL). Abesthood. Obesity is linked to high serum cholesterol levels, high triglyceride levels, high blood pressure, and diabetes. This risk can be reduced by losing just 10% of one's body weight. Diabetes. Diabetes. Inadequate or secreted insulin hormones allow the body to raise blood sugar levels, increasing your risk of a heart attack. Insulin has no effect on the amount of blood sugar you get. Significant metabolism. This is due to your smoking, asthma, and high blood sugar levels. You will have heart disease twice as often if you have metabolic syndrome.

The heart attacks the family's past. You may be at higher risk because your siblings, parents, or ancestors had heart attacks at a young age (55 for men, 65 for women). Inability to exercise. Obesity and inactive serum cholesterol both raise blood levels. People who practice daily have better cardiovascular health, including lower blood pressure. Stress: You should respond to stress in order to increase your chances of having a heart attack. Illegal drug use If relaxing narcotics such as cocaine or amphetamines are used, a coronary artery spasm will result in a heart attack. A history of pre-eclampsia This syndrome increases the risk of lifetime heart failure by causing high blood pressure during pregnancy. An autoimmune disorder. A disease such as rheumatoid arthritis or lupus raises the risk.

As a result, the possibility of a heart attack defends this disease's analysis.

Chapter 2

Literature Reviews

Cardiovascular disease had become the leading cause of death in the United States by 1940 [1]. When President Franklin D. Roosevelt died in 1945 of hypertensive heart disease, Americans became interested in heart disease research [2]. The Framingham Heart Study was founded in the United States in 1948 to study heart disease and its various risk factors [3]. Since then, there has been continuous research on heart failure and its various risk factors, with the goal of better preventing it with modern technology. In his paper, Fizar Ahmed [4] explained the architecture of heart attack rates and used the IoT concept to predict future heart attack patients. He also used one of the most popular machine learning algorithms, kNN (k Nearest Neighbour), to complete his work and achieve better accuracy. Prince Kansal et al. [5] focused on how to predict heart disease early using various data mining techniques, and they used nearly four machine learning algorithms. They used age, gender, blood pressure, and blood sugar levels in their dataset. They got better results from the decision tree. Manually diagnosing heart disease, as Asha Rajkumar [2] points out, takes a long time and requires the assistance of experts. They were particularly interested in her paper on fast diagnosis of heart disease using data mining techniques. Three algorithms and tanagra tools were used. The Naive Bayes algorithm took 609ms to diagnose the heart disease.

Chapter 3

Data Collection & Processing

In this section, we present the dataset we used for our investigation and discuss the final dataset processing methods.

3.1 Dataset Overview

For datasets, I have used dataset from Heart Disease Data Set of UCI Machine Learning Repository [6]. This database contains 76 attributes, but all published experiments use only 14 of them.

The patients' names and social security numbers were recently removed from the database and replaced with dummy values. Among them I have decided to use following attributes:

1. Age
2. Sex
    - 1: Male
    - 0: Female
3. Chest Pain Type
    - 1: Typical Angina
    - 2: Atypical Angina
    - 3: Non-anginal Pain
    - 4: Asymptomatic
4. Resting Blood Pressure (in mm Hg on admission to the hospital)

5. Serum Cholesterol in mg/dl
6. Fasting Blood Sugar > 120 mg/dl
    - 1: True
    - 0: False
7. Resting Electrocardiographic Results
    - 0: normal
    - 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
    - 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8. Maximum Heart Rate Achieved
9. Exercise Induced Angina
    - 1: Yes
    - 0: No
10. ST depression induced by exercise relative to rest
11. The Slope of the Peak Exercise ST Segment
    - 1: Upsloping
    - 2: Flat
    - 3: Downsloping
12. Number of Major Vessels (0-3) Colored by Flourosopy
13. Thal
    - 3 = normal
    - 6 = fixed defect
    - 7 = reversable defect
14. Diagnosis of Heart Disease (Angiographic Disease Status)
    - Value 0: < 50% diameter narrowing
    - Value 1: > 50% diameter narrowing

A sample of the dataset is given below.

	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: Sample Dataset

In total the dataset 303 data, 242 data in train.csv portion and 61 data in test.csv portion and in total the whole merged dataset contains 303 unique data. The overview of the dataset is given below.

	0 Labeled (Less Chance)	**1 Labeled (More Chance)**
Train.csv = 242	110	132
Test.csv = 61	28	33
Total = 303		

Table 3.1: Data Distribution of the dataset

3.2 Data Splitting

I divided the final data for training and testing in an 8:2 ratio. 80% of the data, or 242 labeled data points, was set aside for training, while 20%, or 61 labeled data points, were set aside for testing.

Total Data	**Train Data**	**Test Data**
303 data	242 data	61 data
100% data	80% data	20% data

Table 3.2: Data Splitting

Chapter 4

Methodology

With the help of four supervised models—Random Forest Tree Classifier, K-Nearest Neighbor, Multilayer Perceptron, and AdaBoosting—I have experimented with the diagnosis of heart disease.

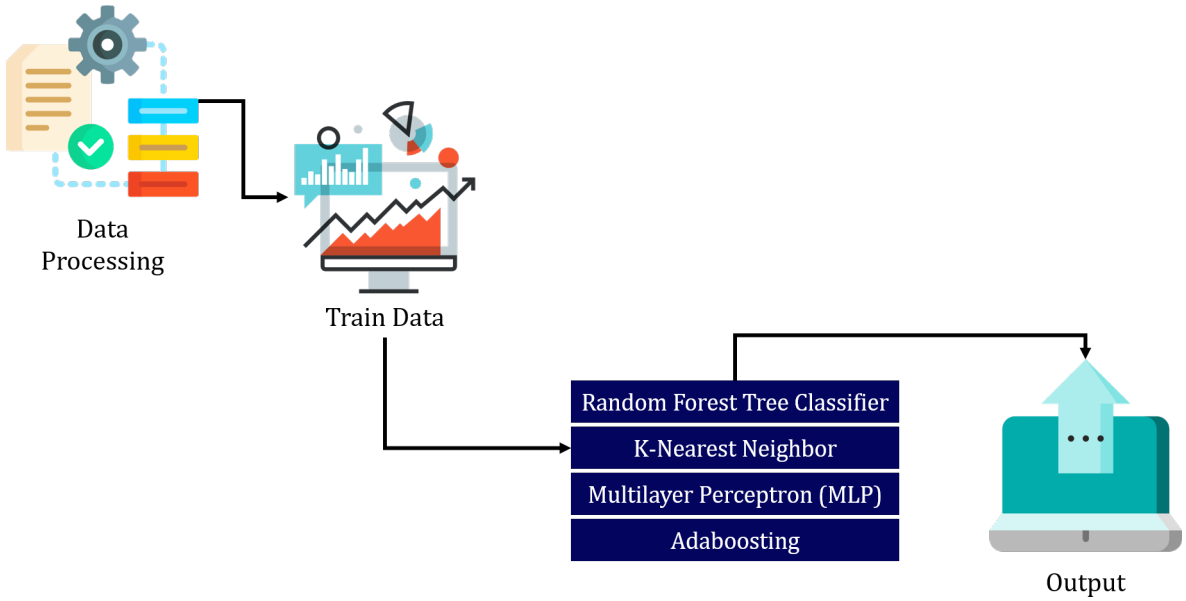

Figure 4.1: Working Methodology

4.1 Random Forest Tree Classifier

4.1.1 About Random Forest

Random forest is a type of Supervised Machine Learning Algorithm that is commonly used in classification and regression problems. It constructs decision trees from various samples

and uses their majority vote for classification and average for regression.

One of the most important characteristics of the Random Forest Algorithm is that it can handle data sets with both continuous and categorical variables, as in regression and classification. It outperforms other algorithms in classification problems.

4.1.2 Steps

These are the steps involved in random forest algorithm:

Step 1: In Random forest, n random records are chosen at random from a data set with k records.

Step 2: For each sample, an individual decision tree is built.

Step 3: Each decision tree will produce a result.

Step 4: For classification and regression, the final output is based on majority voting or averaging.

Consider the fruit basket as an example, as shown in the figure below. Now, n samples are drawn from the fruit basket, and an individual decision tree is built for each sample. As shown in the figure, each decision tree will produce an output. The final result is determined by majority voting. As shown in the figure below, the majority decision tree produces an apple rather than a banana, so the final output is an apple.

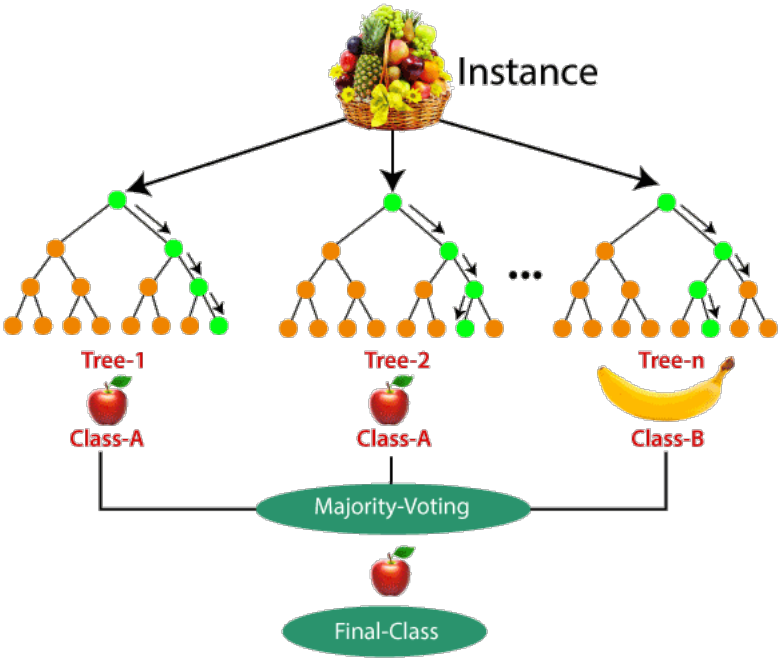

Figure 4.2: Random Forest Example

4.2 Multi layer Perceptron (MLP)

The perceptron is extremely useful for classifying data sets that can be separated linearly. They run into serious problems with data sets that do not follow this pattern, as demonstrated by the XOR problem. The XOR problem demonstrates that there exists a set of four points that are not linearly separable for any classification of four points.

The MultiLayer Perceptron (MLPs) overcomes this limitation by classifying datasets that are not linearly separable. They accomplish this by learning regression and classification models for difficult datasets using a more robust and complex architecture.

4.3 K-Nearest Neighbours

The k-nearest neighbors algorithm, sometimes referred to as KNN or k-NN, is a supervised learning classifier that employs proximity to produce classifications or predictions about the grouping of a single data point. Although it can be applied to classification or regression issues, it is commonly employed as a classification algorithm because it relies on the idea that comparable points can be discovered close to one another.

A class label is chosen for classification problems based on a majority vote, meaning that the label that is most commonly expressed around a particular data point is adopted. Despite the fact that this is officially "plurality voting," literature more frequently refers to "majority vote." The difference between both terms is that "majority voting" informally calls for a majority of more than 50%, which typically only applies when there are only two options. You don't absolutely need 50% of the vote to draw a conclusion about a class when there are many classes, such as four categories; you might assign a class label with a vote of more than 25%.

4.4 Adaboosting

AdaBoost, also known as Adaptive Boosting, is a machine learning method used in an ensemble setting. Decision trees with one level, or Decision trees with only one split, are the most popular algorithm used with AdaBoost. Another name for these trees is Decision Stumps.

This algorithm creates a model while assigning each data piece an equal weight. Then, it gives points that were incorrectly categorised larger weights. The next model now gives more weight to all the points with higher weights. If no lowe error is received, it will continue to train the models.

Chapter 5

Experiments and Results

5.1 K-Nearest Neighbor

5.1.1 Train Average Accuracy

I got **82.21%** accuracy for K-Nearest Neighbor.

5.1.2 Experimental Results

Let's see the Experimental Results of K-Nearest Neighbor:

	precision	**recall**	**f1-score**	**support**
less chance	0.81	0.79	0.80	28
more chance	0.82	0.85	0.84	33
accuracy			0.82	61
macro avg	0.82	0.82	0.82	61
weighted avg	0.82	0.82	0.82	61

Table 5.1: K-Nearest Neighbor Classification Report

5.1.3 Confusion Matrix

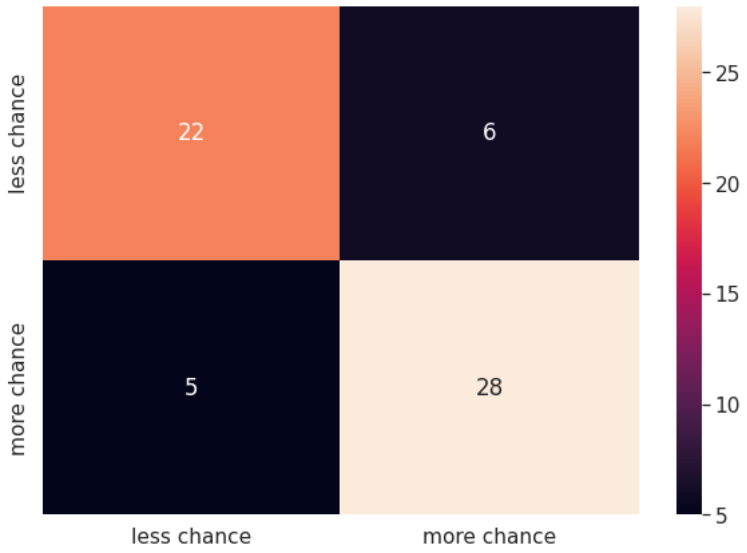

Figure 5.1: Confusion Matrix of K-Nearest Neighbor

5.2 Multilayer Perceptron

5.2.1 Train Average Accuracy

I got **80.97%** accuracy for Multilayer Perceptron.

5.2.2 Experimental Results

Let's see the Experimental Results of Multilayer Perceptron (MLP):

	precision	**recall**	**f1-score**	**support**
less chance	0.72	0.75	0.74	28
more chance	0.78	0.76	0.77	33
accuracy			0.75	61
macro avg	0.75	0.75	0.75	61
weighted avg	0.76	0.75	0.75	61

Table 5.2: Multilayer Perceptron Classification Report

5.2.3 Confusion Matrix

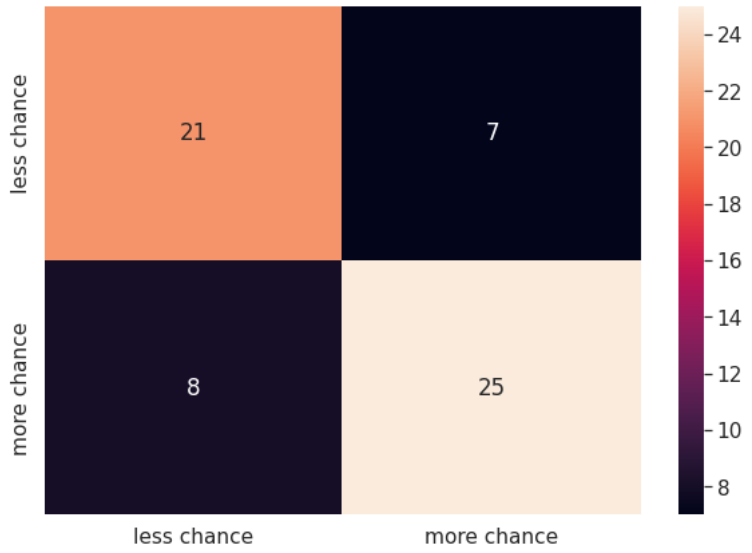

Figure 5.2: Confusion Matrix of K-Nearest Neighbor

5.3 Adaboosting

5.3.1 Train Average Accuracy

I got **82.63%** accuracy for Adaboosting.

5.3.2 Experimental Results

Let's see the Experimental Results of Adaboosting:

	precision	**recall**	**f1-score**	**support**
less chance	0.79	0.79	0.79	28
more chance	0.82	0.82	0.82	33
accuracy			0.80	61
macro avg	0.80	0.80	0.80	61
weighted avg	0.80	0.80	0.80	61

Table 5.3: Adaboosting Classification Report

5.3.3 Confusion Matrix

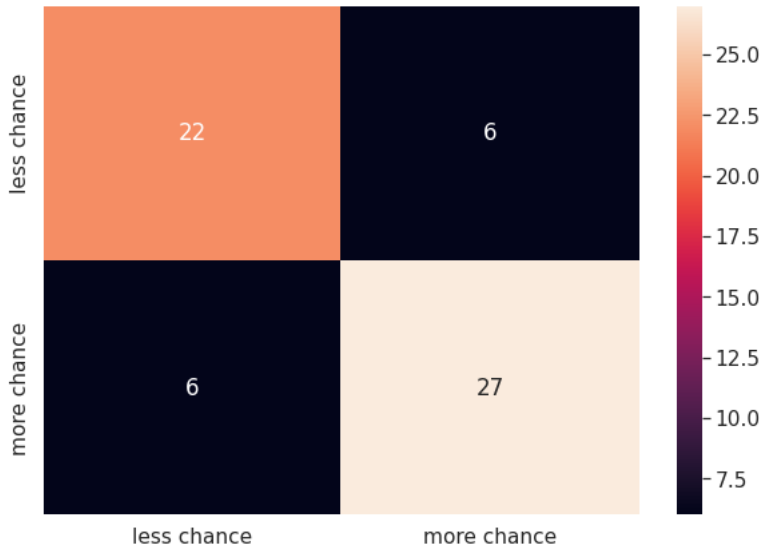

Figure 5.3: Confusion Matrix of Adaboosting

5.4 Random Forest

5.4.1 Train Average Accuracy

I got **84.68%** accuracy for Random Forest Tree Classifier.

5.4.2 Experimental Results

Let's see the Experimental Results of Random Forest Tree:

	precision	**recall**	**f1-score**	**support**
less chance	0.76	0.79	0.77	28
more chance	0.81	0.79	0.80	33
accuracy			0.79	61
macro avg	0.79	0.79	0.79	61
weighted avg	0.79	0.79	0.79	61

Table 5.4: Random Forest Classification Report

5.4.3 Confusion Matrix

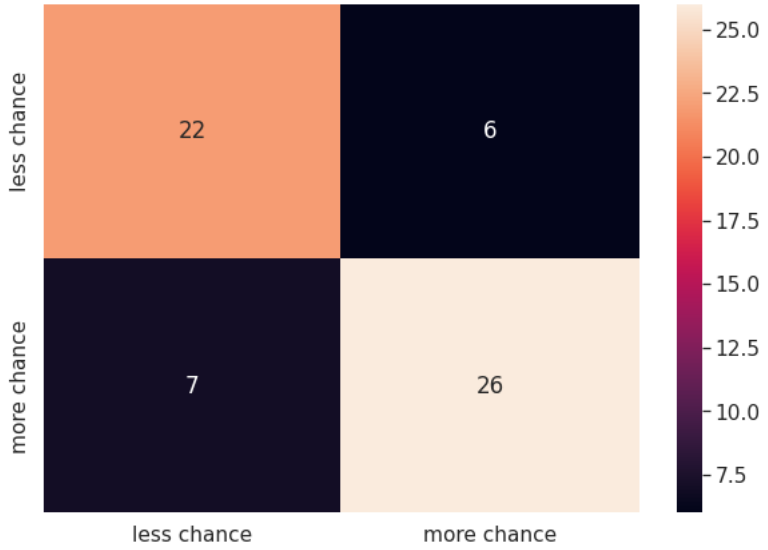

Figure 5.4: Confusion Matrix of Random Forest

5.5 Bar Chart

5.5.1 Accuracy Bar chart for train set

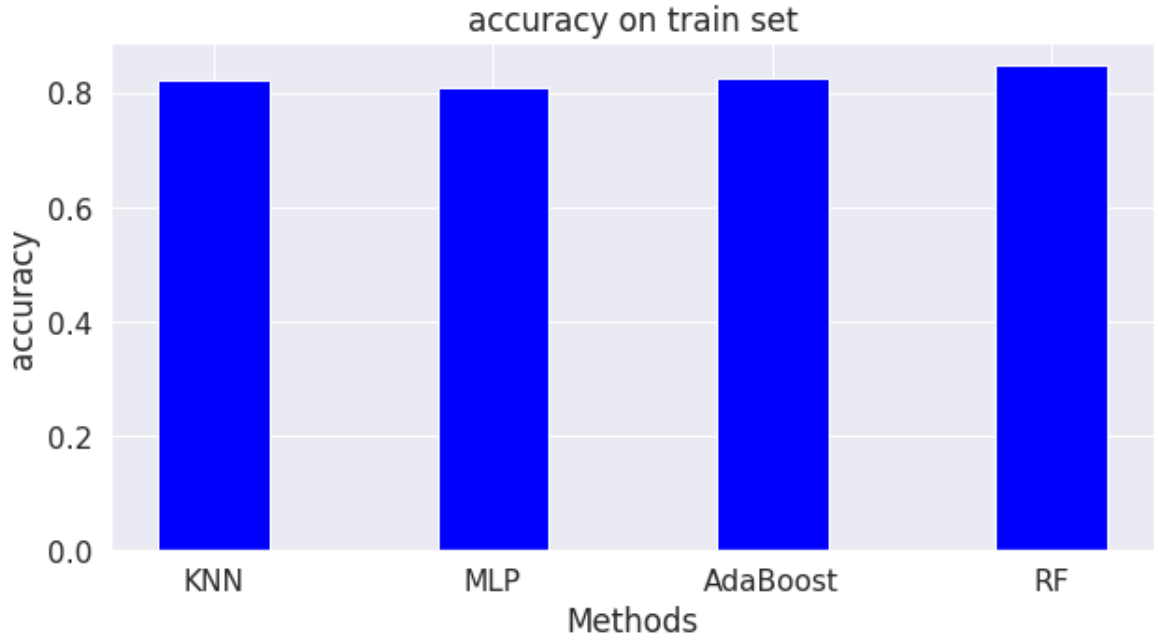

Figure 5.5: Accuracy Bar chart for train set

5.5.2 Accuracy Bar chart for test set

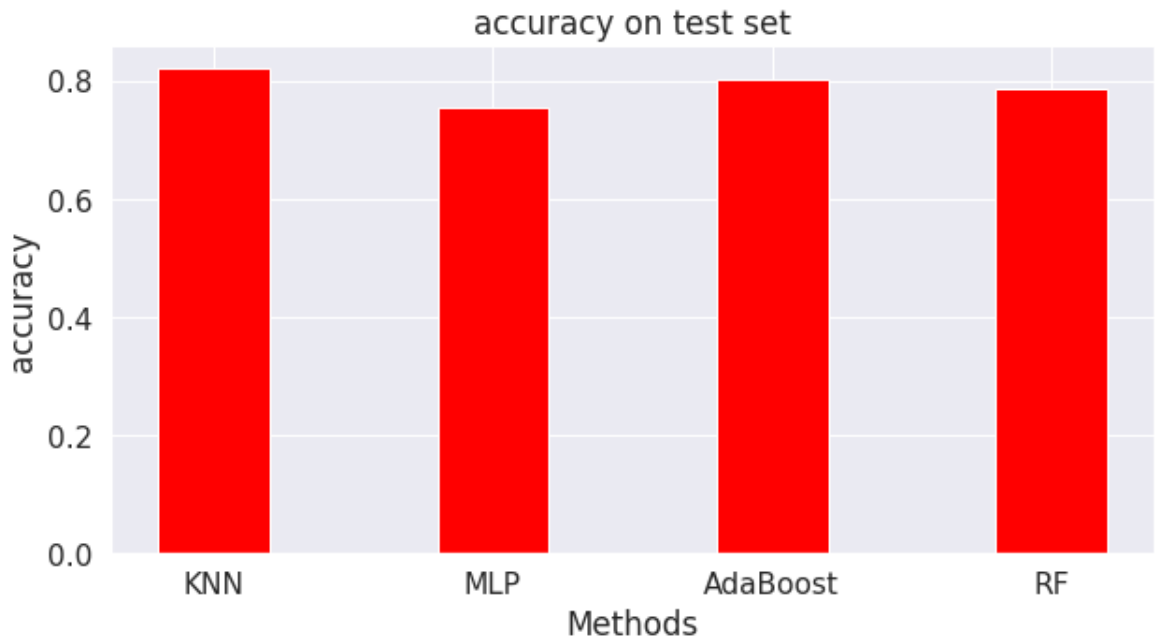

Figure 5.6: Accuracy Bar chart for test set

5.6 Accuracy Comparison of models

Model Name	Training Accuracy	Test Accuracy
KNN	82.2%	82%
MLP	81%	75.4%
AdaBoost	82.6%	80.3%
RF	84.7%	78.7%

Table 5.5: Model Comparison

Here we can see that RF working the best on training data and KNN giving the best output on test dataset.

5.7 ROC curve Analysis

5.7.1 ROC curve for train and test Data

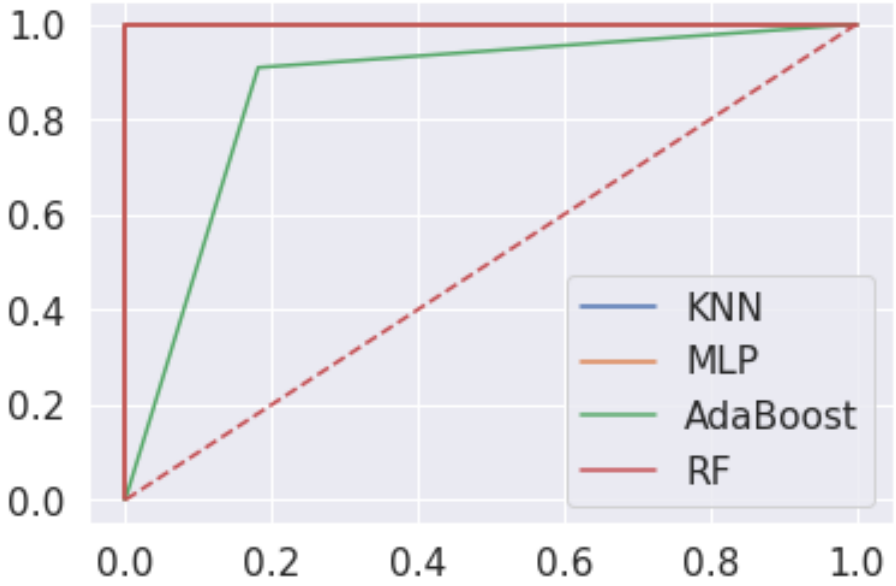

Figure 5.7: ROC for train data

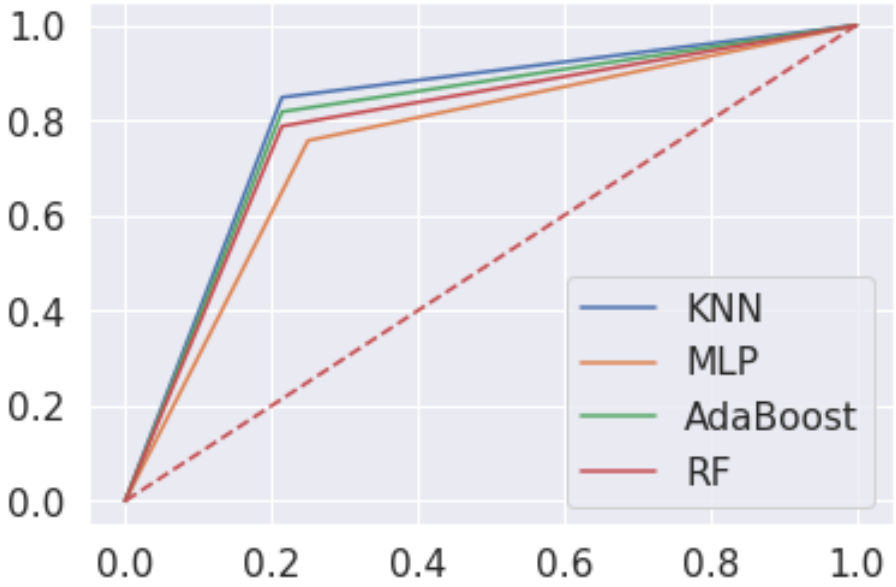

Figure 5.8: ROC for test data

The ROC curve stands for Receiver Operating Characteristic curve. ROC curves display the performance of a classification model.In terms of expected probability, ROC shows us how successful the model is at differentiating between the provided classes. The greater the AUC for the ROC Curve, the better the performance. We can see from the constructed curves that the KNN has the highest AUC, indicating that it is the best model for this classifier.

Chapter 6

Future Work and Conclusion

6.1 Future Works

My experiment produced good results, but the accuracy may be increased by testing with more data and more sophisticated ways. Using several models can also assist to enhance performance.

6.2 Conclusion

The heart attack possibility performance is shown in this research analysis. The performance analysis identified several categories, including confusion metrics, precision, recall, f measure, and auc. In terms of overall performance, KNN outperforms the others, with an accuracy of 82%. To determine the best performance from our dataset, some machine learning techniques are used. In the future, this study will incorporate deep learning and artificial intelligence methods to analyze and predict the possibility of a heart attack. I also added more information as needed.

References

[1] W. B. Kannel, "Contribution of the framingham study to preventive cardiology," *Journal of the American College of Cardiology*, vol. 15, no. 1, pp. 206–211, 1990.

[2] A. Rajkumar and G. S. Reena, "Diagnosis of heart disease using datamining algorithm," *Global journal of computer science and technology*, vol. 10, no. 10, pp. 38–43, 2010.

[3] S. S. Mahmood, D. Levy, R. S. Vasan, and T. J. Wang, "The framingham heart study and the epidemiology of cardiovascular disease: a historical perspective," *The lancet*, vol. 383, no. 9921, pp. 999–1008, 2014.

[4] F. Ahmed, "An internet of things (iot) application for predicting the quantity of future heart attack patients," *International Journal of Computer Applications*, vol. 164, no. 6, pp. 36–40, 2017.

[5] A. Methaila, P. Kansal, H. Arya, P. Kumar, *et al.*, "Early heart disease prediction using data mining techniques," *Computer Science & Information Technology Journal*, vol. 28, pp. 53–59, 2014.

[6] "Heart disease data set." https://archive.ics.uci.edu/ml/datasets/Heart+Disease. Accessed: 2022-08-15.