Heart Attack Analysis and Prediction

Mir Adis Ali (182014015)[1], Mahisur Rahman (182014016)[2], Md. Shihan (193014059)[3]

Department of Computer Science and Engineering
[1] adis.ali.cse@ulab.edu.bd
[2]mahisur.rahman.cse@ulab.edu.bd
[3]md.shihan.cse@ulab.edu.bd

University of Liberal Arts Bangladesh

Abstract

Heart Attack Analysis and Prediction using Machine Learning Technique in Big data analytics have started to play a vital role in healthcare practices and research. coronary failure prediction will be found totally on real-time operation, distributed and real-time classification and distribution, and storage so; databases are often easily modified by the doctors. If you recognize all the attributes associated with our health we can check easily what quantity chance of the center attack risk, using the system applications. it had been recently accustomed train classification and models. After that using extract the features that are a condition to be found to be classified by the Decision Tree, Logistic Regression, Support Vector Machine and Random Forest. Compared to existing; algorithms provide better performance. After classification, performance criteria including accuracy, precision, and F-measure is to be calculated. If you are concerned about the guts attack risks, you would possibly be mentioned by a cardiologist. Some attributes are coronary failure risk factors which are High pressure, high cholesterol and diabetes, which increase your risk even more. Hence, we also are checking your symptoms of coronary failure and talking about prevention.

Introduction

The heart may be a muscle and its role is to pump blood throughout the body. This makes the body a serious staple. Heart the disease is one in all the largest health risks for association today. per the World Health Organization (WHO), stroke and heart attacks are the foremost common cause of global death (85%). Therefore, the availability of knowledge and data processing techniques, especially machine learning and early detection of Heart Attacks, can help patients anticipate a possible disease response. within the healthcare field, it is becoming more and more common nowadays to source large amounts of knowledge (big data), streaming machines, advanced

healthcare services, high throughput instruments, sensor networks, Internet of Things, mobile application applications, data archiving and processing, from many areas.

Related Work

Previous research studies have examined the use of machine learning techniques to predict and classify cardiopathy. However, these studies specialize in the specific effects of a specific machine learning techniques. This work analyses the predictive system for cardiopathy. during this work, medical terms like sex, vital sign, and cholesterol are will not describe the possibility of cardiopathy in patients with 12 points. So far, 13 attributes are used for forecasting. Two more points are added to the present research work - obesity and smoking. data processing classification algorithms, decision trees, navy bias and neural networks are analyzed within the cardiovascular database [1].

Medical diagnostic systems play an important role in practice and are used by medical professionals for diagnosis and treatment. during this work, the medical diagnostic system is defined to indicate the chance of disorder. The system is constructed by combining the relative advantages of genetic mechanisms and neural networks. Multi-layered feed forward neural networks typically adapt to complex classification problems. The weight of the nerve space is set using the genetic technique because it finds an honest set of excellent weights at low repetitions [2].

The condition of the center is explained in detail by a radical examination of the features of the Electrocardiogram report. It is valuable to automatically remove the features of the time plane to detect essential cardiovascular disease. This function introduces a multi-resolution wavelet transform-based system to detect 'P', 'Q', 'R', 'S', 'T' peaks complexes from the original ECG signal [3] .

Dataset and Features

The dataset is publicly available on the Kaggle Website at [4] which is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. It provides patient information which includes over 300 records and 10 attributes. The attributes include: age, sex, chest pain, resting blood pressure, cholestoral, fasting blood sugar, resting electrocardiographic, maximum heart rate, exercise induced angina and Previous peak. The data set is in csv (Comma Separated Value) format which is further prepared to data frame as supported by pandas library in python.

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
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
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

302 rows × 14 columns

Fig.1: Dataset

The education data is irrelevant to the heart disease of an individual, so it is dropped. Further with this dataset pre-processing and experiments are then carried out.

Methods

During this project, we have tried 5 algorithms for experiments and they are Decision Tree, Logistic Regression, Support Vector Machine and Random Forest:

Decision Tree:

Decision Tree may be a simple and simple to implement classifier. The bit through feature to access thorough patients' profiles is simply obtainable in Decision Trees. A decision tree builds classification or regression models within the structure of a tree making it simple to debug and handle. Decision trees can handle both categorical and numerical data. The algorithm works by finding the data gain of the attributes and putting off the attributes for splitting the branches into threes. the knowledge gained for the tree is identified using the below given

$$E(S) = -P(P)log_2P(P) - P(N)log_2P(N)$$

The algorithm for the choice tree is given below:

Step 1: Identify the data gain for the attributes within the dataset.

Step 2: Sort the data gain for the guts disease datasets in descending order.

Step 3: After the identification of the data gain assign the simplest attribute of the dataset at the foundation of the tree.

Step 4: Then calculate the knowledge gained using the identical formula.

Step 5: Split the nodes that supported the very best information gain value.

Step 6: Repeat the method until each attribute is set as leaf nodes all told the branches of the tree.

Logistic Regression:

Logistic Regression may be supervised learning that computes the chances for classification problems with two outcomes. It can even be extended to predict several classes. In the Logistic Regression model, we apply the sigmoid function, which is

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

This function successfully maps any number into the worth between 0 and 1 and we can regard this value because of the probability of predicting classes. as an example, we have got two classes and they are the presence of heart condition and the absence of disease. If we set the edge as 0.5, applying the sigmoid function gives us a price of 0.7, which suggests the person has a 70% probability of getting cardiovascular disease so we will predict that he has cardiopathy.

Support Vector Machine:

SVM aims to find a hyperplane in multiple dimensions (multiple features) that classifies the dataset. Here is a picture of classification by SVM.

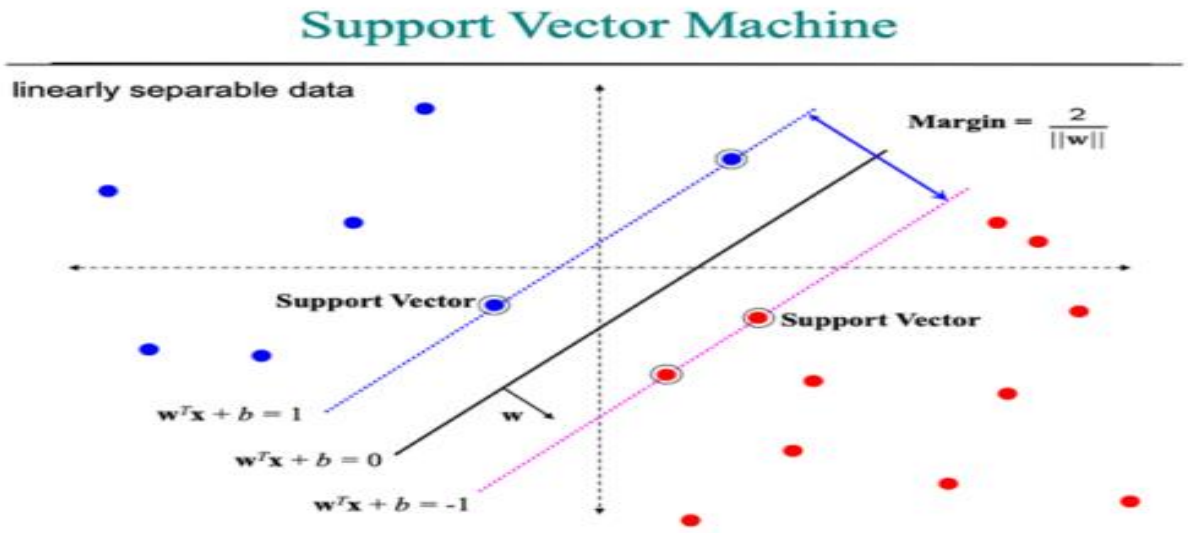

Fig.2: Classification by SVM

Random Forest:

Random Forest is an ensemble learning method for classification and regression by constructing multiple decision trees in training and outputting the classification or prediction(regression). The goal of Random Forest is to mix weak learning models into robust and robust learning models. From an online, we learn that the algorithm of Random Forest will be summarized in 4 steps:

Step 1: Randomly draw M bootstrap samples from the training set with replacement.

Step 2: Grow a choice tree from the bootstrap samples. At each node: Randomly select K features without replacement and split the node by finding the simplest cut among the chosen features that maximize the data gain.

Step 3: Repeat the steps 1 and a pair of T times to urge T trees;

Step 4: Aggregate the predictions made by different trees via the bulk vote.

Experiments/Results/Discussion

Exploratory Analysis: Correlation Matrix visualization

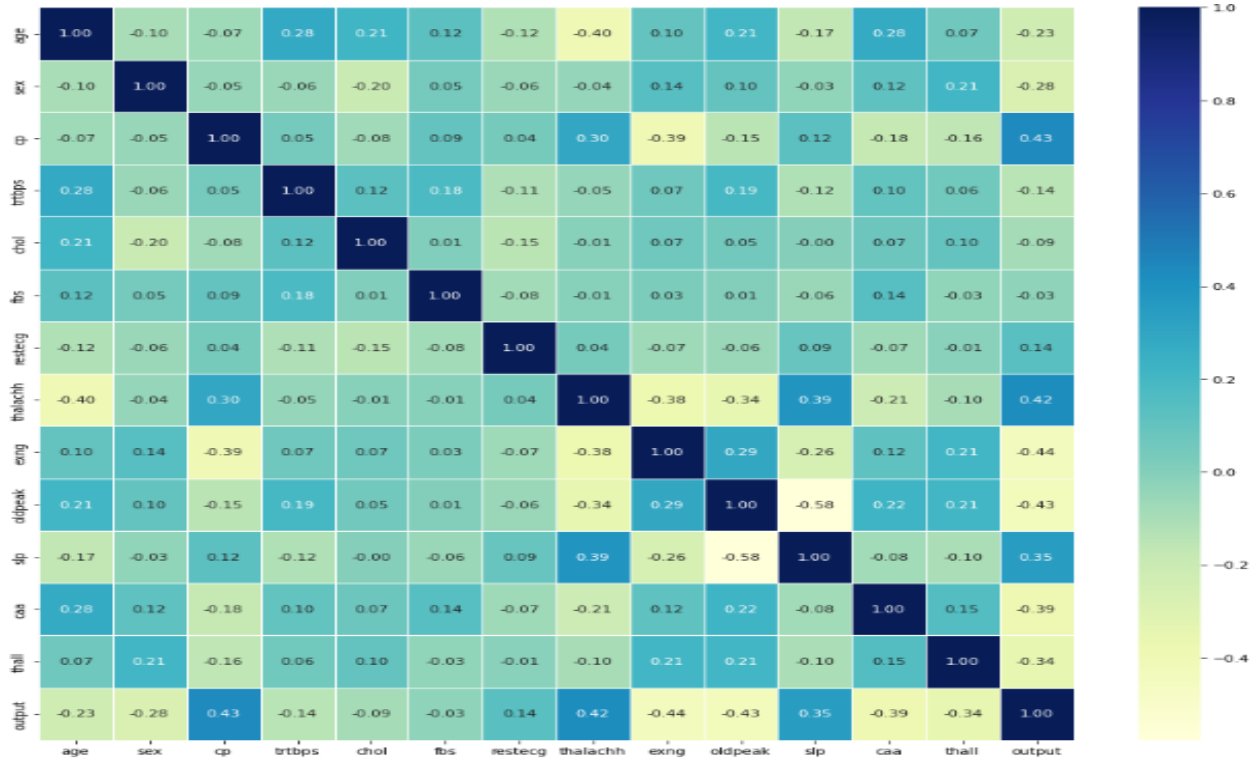

Fig.3: Correlation Matrix Visualization

It shows that there is no single feature that features a very high correlation with our target value. Also, a number of the features correlate with the target value and a few have a positive. the information was also visualized through plots and bar graphs.

Since our project could be a classification problem, we use test accuracy, cross-validation accuracy, precision, recall and F1 to gauge the models. Here is the table of results of various methods and we will discuss each evaluation of methods in detail.

Methods	Test Accuracy	Cross-Validation Accuracy	AUC	Precision	Recall	F1
Decision Tree	80.3	0.68	0.81	0.92	0.78	0.84
Logistic Regression	88.5	0.77	0.92	0.87	0.90	0.89
SVM	86.9	0.41	0.84	0.87	0.87	0.87
Random Forest	83.6	0.84	0.94	0.84	0.84	0.84

Decision Tree:

The confusion matrix of Decision Tree is

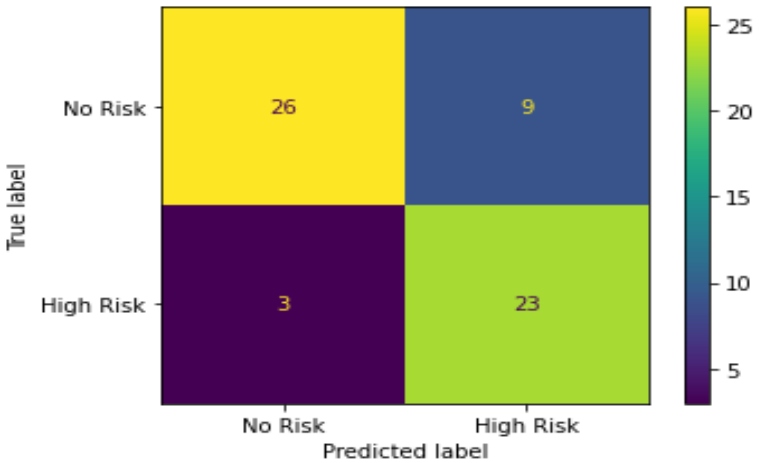

Fig.4: Confusion Matrix for Decision Tree

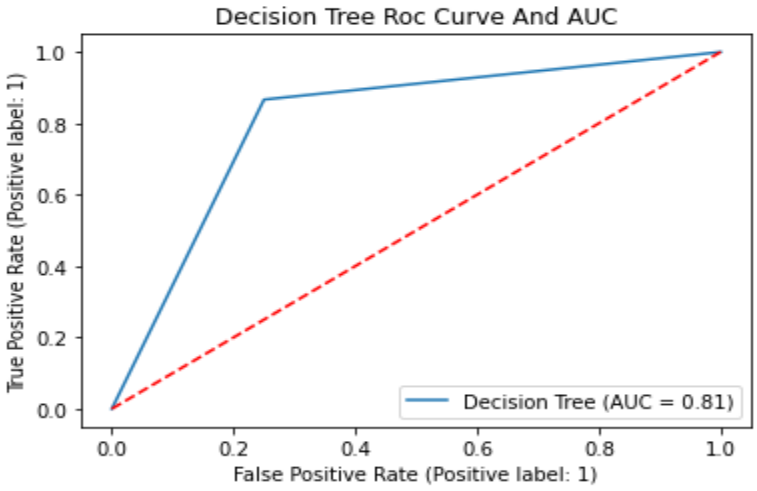

Fig.5: Decision Tree Roc Curve and AUC

Logistic Regression:

Here is the confusion matrix of the Logistic Regression

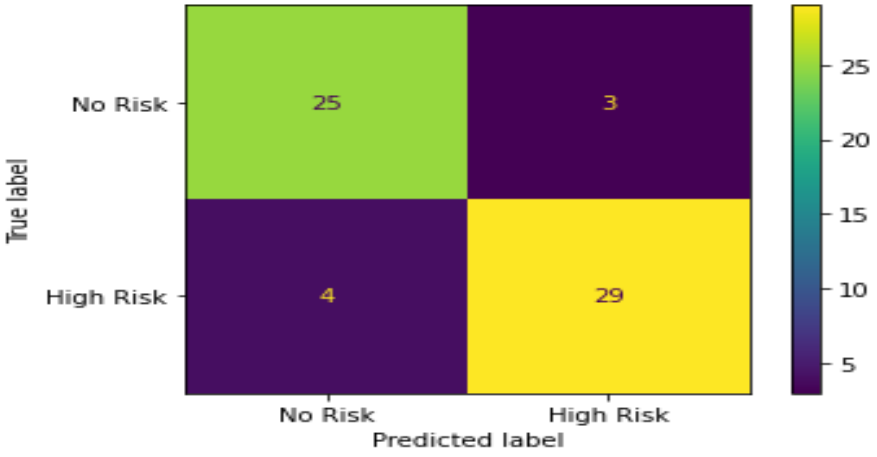

Fig.6: Confusion Matrix for Logistic Regression

The square of the magnitude of coefficients is supported by Logistic Regression to avoid overfitting. The test accuracy is 88.5%. It performs well but is not the most effective for us. The advantage of the Logistic Regression is that it does not need an excessive amount of computational resources and it is highly interpretable. So, it is easy and sufficient to use Logistic Regression. However, the limitation of Logistic Regression is that it assumes linearity between the features of the dataset. Within the world, the information is never separable, nor as our dataset. that is why we cannot reach a really high accuracy of 90%.

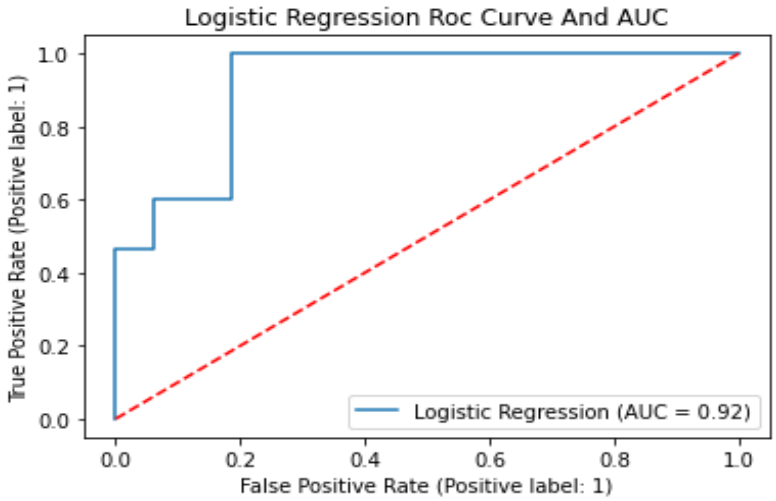

Fig.7: Logistic Regression Roc Curve and AUC

SVM:

Here is the confusion matrix for SVM

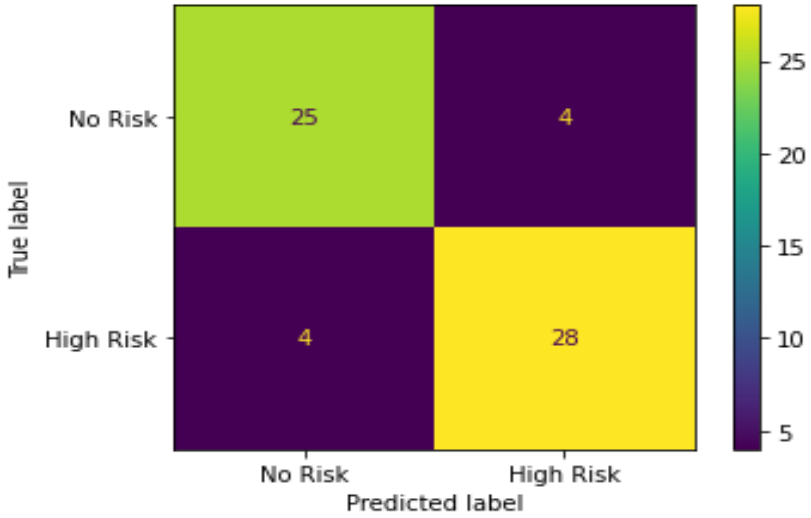

Fig.8: Confusion Matrix for SVM

According to the tutorial of sklearn, for atiny low dataset, it is better to use sklearn. svm.SVC(). The test accuracy is 86.9%. The advantage of SVM is that it is very efficient with high dimensional spaces. The most disadvantage is that the SVM has many parameters that have to be correctly chosen to attain the most effective performance. For safety, we just use the default parameters of SVM. and therefore the test accuracy of 86.9%.

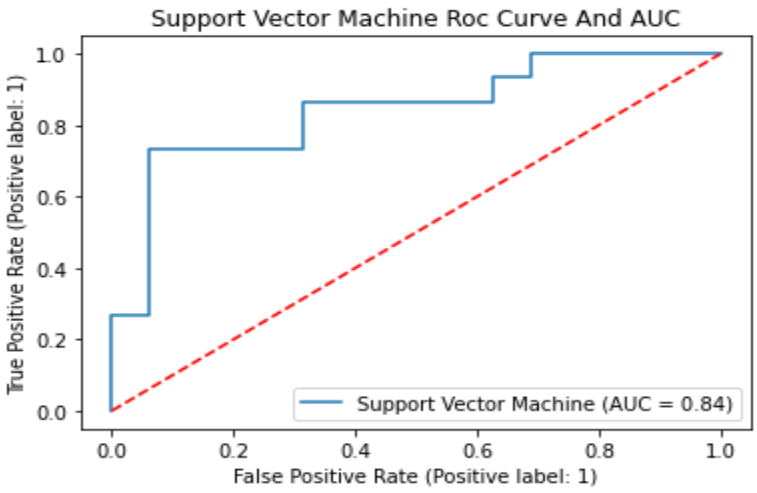

Fig.9: Support Vector Machine Roc Curve and AUC

Random Forest:

The confusion matrix of Random Forest is

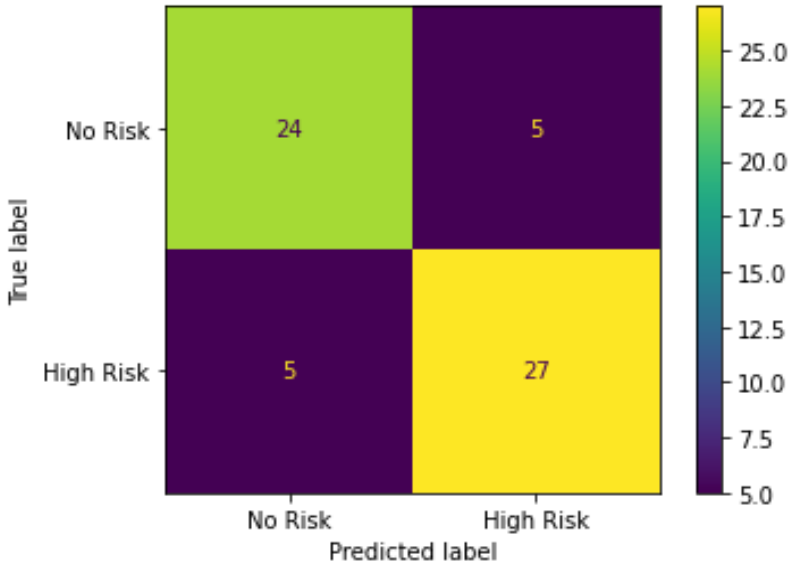

Fig.10: Confusion Matrix for Random Forest

The test accuracy is 83.6%. At the primary beginning, we use the default parameters (n_estimators=100, which suggests the number of trees within the forest is 100 and max_depth = None, which suggests the nodes are expanded until all leaves are pure or all leaves contain but the minimum number of samples required to separate an inside node). We guess it had to be overfitting. One reason may be the training data is not generalized during the training process so we arrange to shuffle the dataset again and that we tried the parameter random_state from 1 to

2000. When random_state is 1826, the test accuracy is 83.6%. Then we tried experiments on parameters of n_estimatros(from 10 to 300) and max_depth(from 10 to 300) and therefore the best test accuracy remains 83.6%. this implies with random_state =1825, the opposite default parameters are adequate to urge the simplest test accuracy. for instance, the amount of trees within the forest is 100, which is acceptable. If the number of trees is little, it will cause underfitting because the model has not been optimized for the training data, plus the test data. If the quantity of trees is just too big, it will cause overfitting because the model becomes so complex and sensitive to new data. The advantage of Random Forest is that it can accommodate datasets with high features and balance the variance and it is not sensitive to the noise of the info. Among these 5 models, Random Forest outperforms the other models.

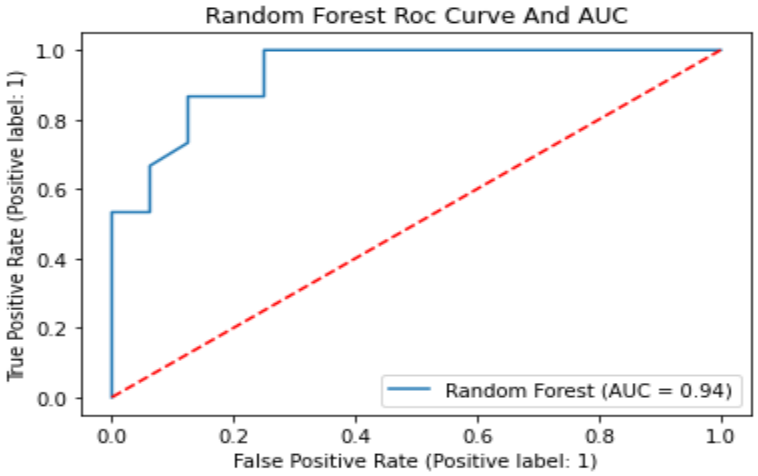

Fig.11: Random Forest Roc Curve and AUC

Conclusion/Future Work

We use some libraries [5] provided by Python to implement this project. After the experiments, the algorithm of Logistic Regression gives us the best test accuracy, which is 88.5%. the rationale why it outperforms others is that it is not limited to the property of the dataset. Random Forest requires the features to be mutually independent. Decision Tree requires the features to be linearly separable. SVM requires the parameters to be appropriately set. Though we get good results of 88.5% accuracy, that is not enough because it cannot guarantee that no wrong diagnosis happens. to enhance accuracy, we hope to need more datasets because 300 instances of dataset do not seem to be sufficient to try and do a superb job. In the future, to predict disease we wish to do different diseases like carcinoma by using image detection. In this way, the dataset becomes complicated and we can apply a convolutional neural network to form accurate predictions.

References

[1] M. B. Youness Khourdifi, "Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization," *International Journal of Intelligent Engineering and Systems 12(1).,* 2019.

[2] K. M. Abderrahmane Ed-Daoudy, "Real-time machine learning for early detection of heart disease using big data approach," *International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS),* 2019.

[3] B. P. M. Akhiljabbar, "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm," 2013.

[4] R. RAHMAN, "Heart Attack Analysis & Prediction Dataset," 2021.

[5] T. Simmons, "10 Essential Data Science Packages for Python," 2019.