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

Report

on

Prediction of Heart Disease using Machine Learning

carried out as part of the course: DS2231

Submitted by

Deivyansh Singh (219309064)

DSE-(IV)B

in partial fulfilment for the award of the degree

of

BACHELOR OF TECHNOLOGY

In

Data Science & Engineering (DSE)

Department of Information Technology (IT)
School of Computing & Information Technology (SCIT)
Manipal University Jaipur,
April 2023

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

Acknowledgement

This project would not have completed without the help, support, comments, advice, cooperation and coordination of various people. However, it is impossible to thank everyone individually; I am hereby making a humble effort to thank some of them.

I acknowledge and express my deepest sense of gratitude of my internal supervisor **Ms. Shipra Shukla** for his/her constant support, guidance, and continuous engagement. I highly appreciate her technical comments, suggestions, and criticism during the progress of this project "**Prediction Of Heart Disease Using Machine Learning**".

I owe my profound gratitude to **Dr. Akhilesh Sharma**, Head of Department (HoD) of DSE, for his valuable guidance and facilitating me during my work. I am also very grateful to all the faculty members and staff for their precious support and cooperation during the development of this project.

Finally, I extend my heartfelt appreciation to my classmates for their help and encouragement.

Student Name:- Deivyansh Singh **Registration No.:- 219309064**

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

Department of Information Technology

School of Computing & Information Technology

Date: _____

CERTIFICATE

This is to certify that the project entitled ***"Prediction Of Heart Disease Using Machine Learning"*** is a bonafide work carried out as ***Project Based Learning (Course Code: DS2231)*** in partial fulfilment for the award of the degree of Bachelor of Technology in CSE-AIML, under my guidance by ***Deivyansh Singh*** bearing registration number ***219309064***, during the academic semester *IV of year 2022-23.*

Place: Manipal University Jaipur, Jaipur

Name of the project guide: _____

Signature of the project guide: _____

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

Contents

Page No.

1. Introduction

1.1 ***Motivation***

Our project object is to detect whether patients have heart disease or not by given a number of features from patients. The motivation of our project is to save human resources in medical centres and improve accuracy of diagnosis. In our project we use different methods to detect heart disease such as Decision Tree, KNN Classifier and Random Forest. And among all these algorithms KNN Classifier gives us the best accuracy of 84.48%.

1.2 ***Abstract***

According to the World Health Organization, every year 12 million deaths occur worldwide due to Heart Disease. The load of cardiovascular disease is rapidly increasing all over the world from the past few years. Many researches have been conducted in attempt to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications. This project aims to predict future Heart Disease by analysing data of patients which classifies whether they have heart disease or not using machine-learning algorithms.

1.3 ***Problem Statement***

The major challenge in heart disease is its detection. There are instruments available which can predict heart disease but either they are expensive or are not efficient to calculate chance of heart disease in human. Early detection of cardiac diseases can decrease the mortality rate and overall complications. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience, time and expertise. Since we have a good amount of data in today's world, we can use various machine learning algorithms to analyse the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data.

2. Literature Review

2.1 Literature Study

Machine Learning techniques are used to analyse and predict the medical data information resources. Diagnosis of heart disease is a significant and tedious task in medicine. The term Heart disease encompasses the various diseases that affect the heart. The exposure of heart disease from various factors or symptom is an issue which is not complimentary from false presumptions often accompanied by unpredictable effects. The data classification is based on Supervised Machine Learning algorithm which results in better accuracy. Here we are using the Random Forest as the training algorithm to train the heart disease dataset and to predict the heart disease. The results showed that the medicinal prescription and designed prediction system is capable of prophesying the heart attack successfully.

Machine Learning techniques are used to indicate the early mortality by analysing the heart disease patients and their clinical records which have brought about the two Machine Learning techniques, k- nearest neighbour model and existing Random Forest to predict the stroke severity index of the patients. This study shows that k-nearest neighbour performed better than Random Forest model. The main objective is to evaluate the different classification techniques, Decision Tree, KNN and Random Forest. After this, evaluating some performance in measures of accuracy, precision, sensitivity, specificity are evaluated.

2.2 Dataset & Features

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

• age: age in years

• sex: sex (1 = male; 0 = female)

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

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

• chol: serum cholesterol in mg/dl

• fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false) • target: Heart disease (0 = no, 1 = yes)

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

3. **Methodology**

3.1 KNN

K-nearest neighbors (KNN) algorithm uses 'feature similarity' to predict the values of new datapoints which further means that the new data point will be assigned a value based on how closely it matches the points in the training set. We can understand its working with the help of following steps −
Step 1 − For implementing any algorithm, we need dataset. So during the first step of KNN, we must load the training as well as test data.

Step 2 − Next, we need to choose the value of K i.e. the nearest data points. K can be any integer.

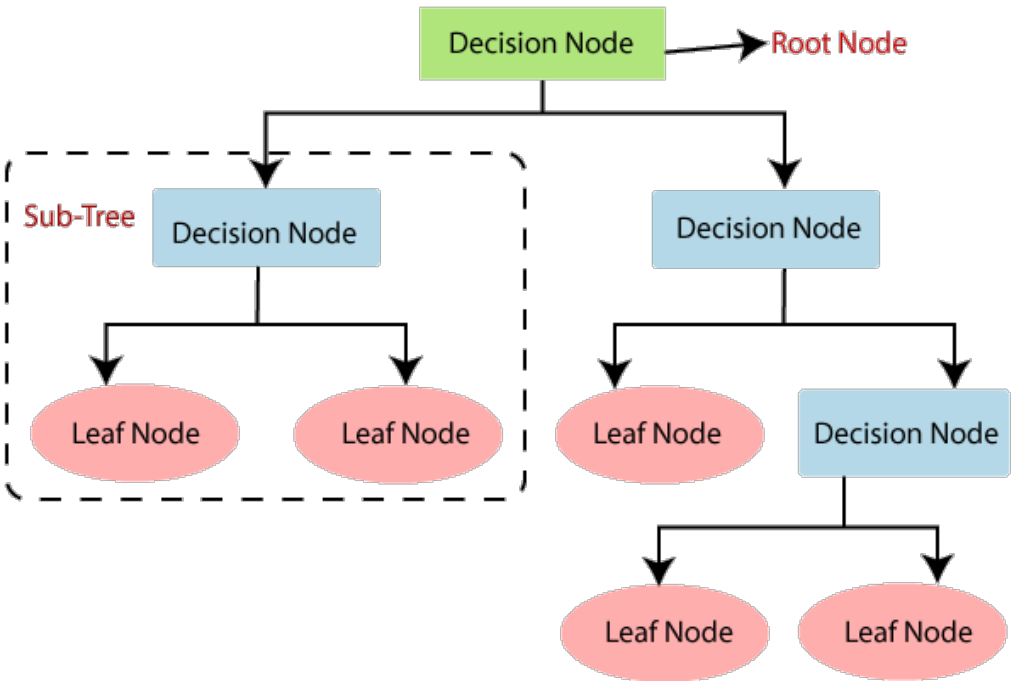
Step 3 − For each point in the test data do the following −

- **3.1** − Calculate the distance between test data and each row of training data with the help of any of the method namely: Euclidean, Manhattan or Hamming distance. The most commonly used method to calculate distance is Euclidean.
- **3.2** − Now, based on the distance value, sort them in ascending order.
- **3.3** − Next, it will choose the top K rows from the sorted array.
- **3.4** − Now, it will assign a class to the test point based on most frequent class of these rows

Step 4 – End

3.2 Decision Tree

- Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.
- In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.



- **Step-1:** Begin the tree with the root node, says S, which contains the complete dataset.
- **Step-2:** Find the best attribute in the dataset using Attribute Selection Measure (ASM)**.**
- **Step-3:** Divide the S into subsets that contains possible values for the best attributes.
- **Step-4:** Generate the decision tree node, which contains the best attribute.
- **Step-5:** Recursively make new decision trees using the subsets of the dataset created in step-3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

3.3 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 combine weak leaning models into a strong and robust leaning model. We learn that the algorithm of Random Forest can be summarized in 4 steps

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

●Step 2: Grow a decision tree from the bootstrap samples. At each node: Randomly select K features without replacement and split the node by finding the best cut among the selected features that maximizes the information gain.

●Step 3:Repeat the steps 1 and 2 T times to get T trees.

●Step 4:Aggregate the predictions made by different trees via the majority vote

3.4 Data Pre-processing

This file contains all the pre-processing functions needed to process all input documents andtexts. First we read the train, test and validation data files then performed some preprocessing like tokenizing, stemming etc. There are some exploratory data analysis is performed like response variable distribution and data quality checks like null or missing values etc.

3.5 Classification

Here we have built all the classifiers for the breast cancer diseases detection. The extracted features are fed into different classifiers. We have used KNN Imputer, Decision Tree and Random forest classifiers from sklearn. Each of the extracted features was used in all of the classifiers. Once fitting the model, we compared the f1 score and checked the confusion matrix. After fitting all the classifiers, 2 best performing models were selected as candidate models for heart diseases classification. We have performed parameter tuning by implementing Grid SearchCV methods on these candidate models and chosen best performing parameters for these classifier. Finally selected model was used for heart disease detection with the probability of truth.

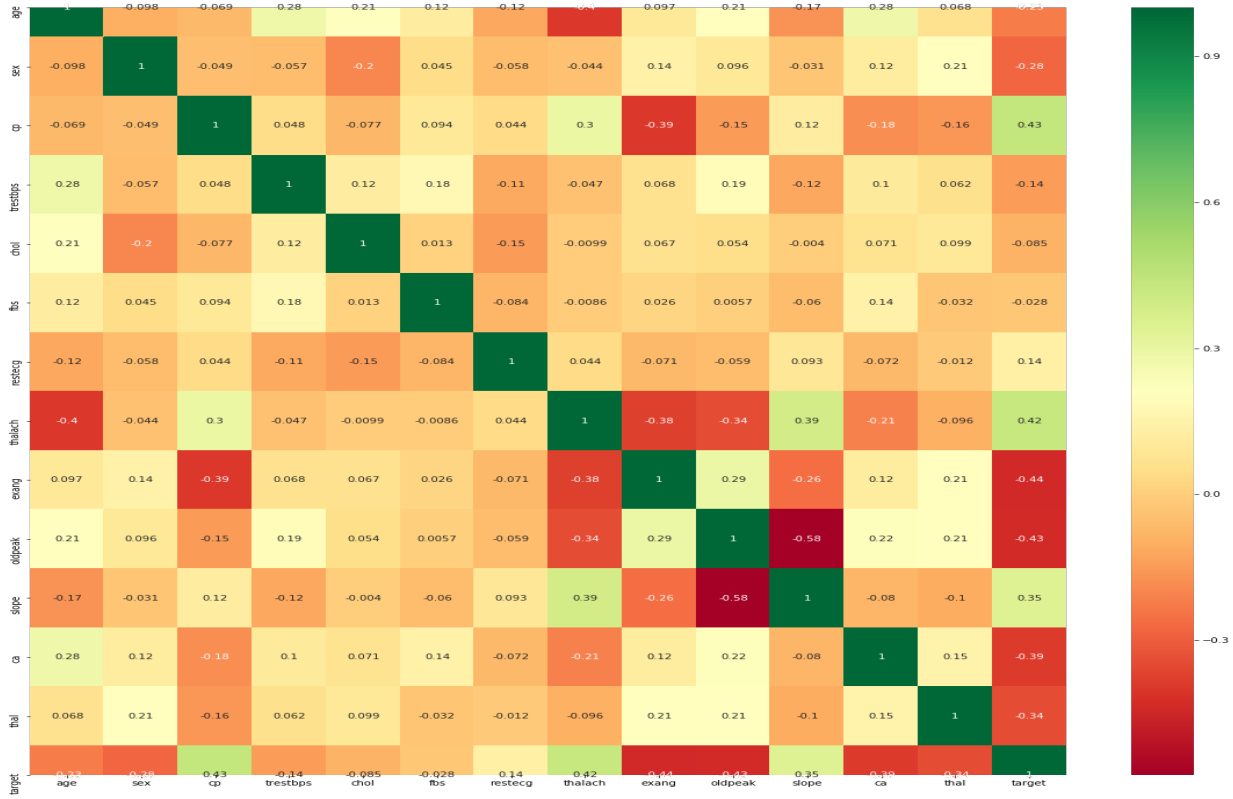
4. **RESULTS**

Since our project is a classification problem, we use accuracy, precision, recall and F1 score to evaluate the models. We would like to introduce the meaning of TP,FP, TN and FN. A true positive (TP) is a positive outcome predicted by the model correctly while a false positive (FP) is a positive outcome predicted by the model incorrectly. A true negative (TN) is a negative outcome predicted by the model correctly while a false negative (FN) is a negative outcome predicted by the model incorrectly. We did not use cross-validation because our dataset is not very sufficient. We split the dataset into 80% for training and 20% for test. Here is the table of results of different methods and we will talk about each evaluation of methods in details

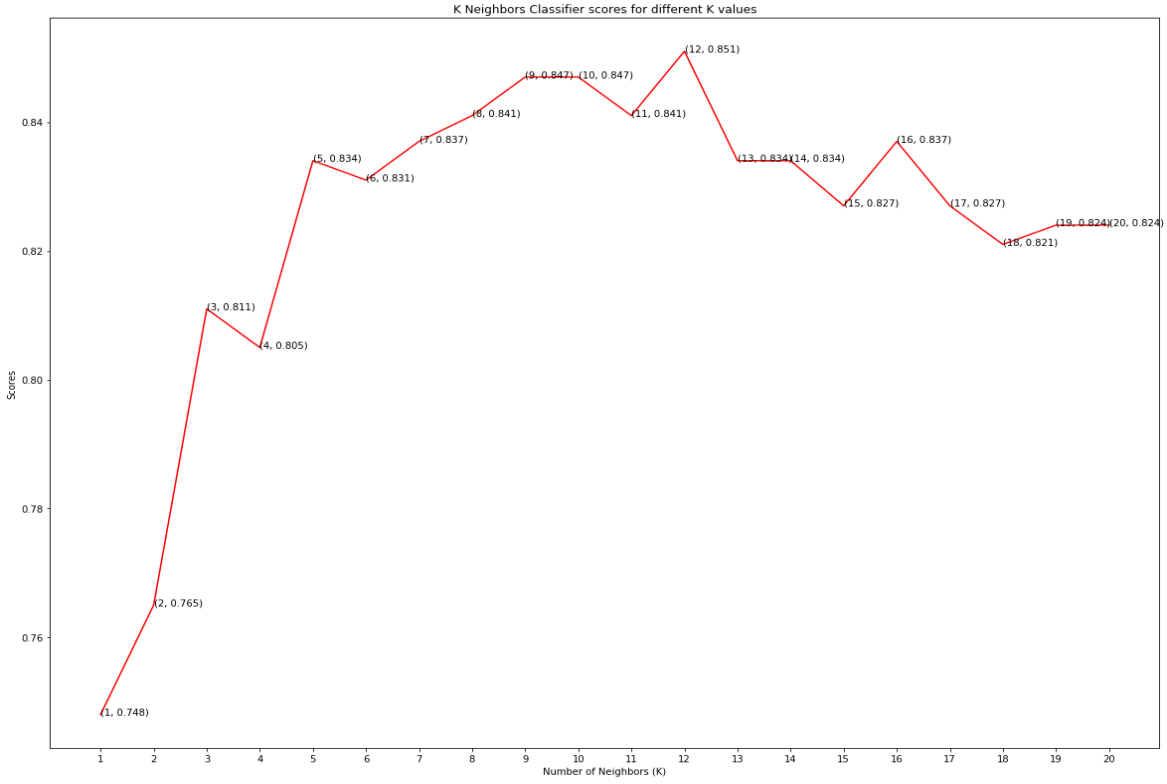
Methods	**Train accuracy**
K Nearest Neighbours	84.48%
Decision Tree	78.51%
Random Forest	81.45%

Here are the results we got after we implemented the following algorithms to the given dataset:-

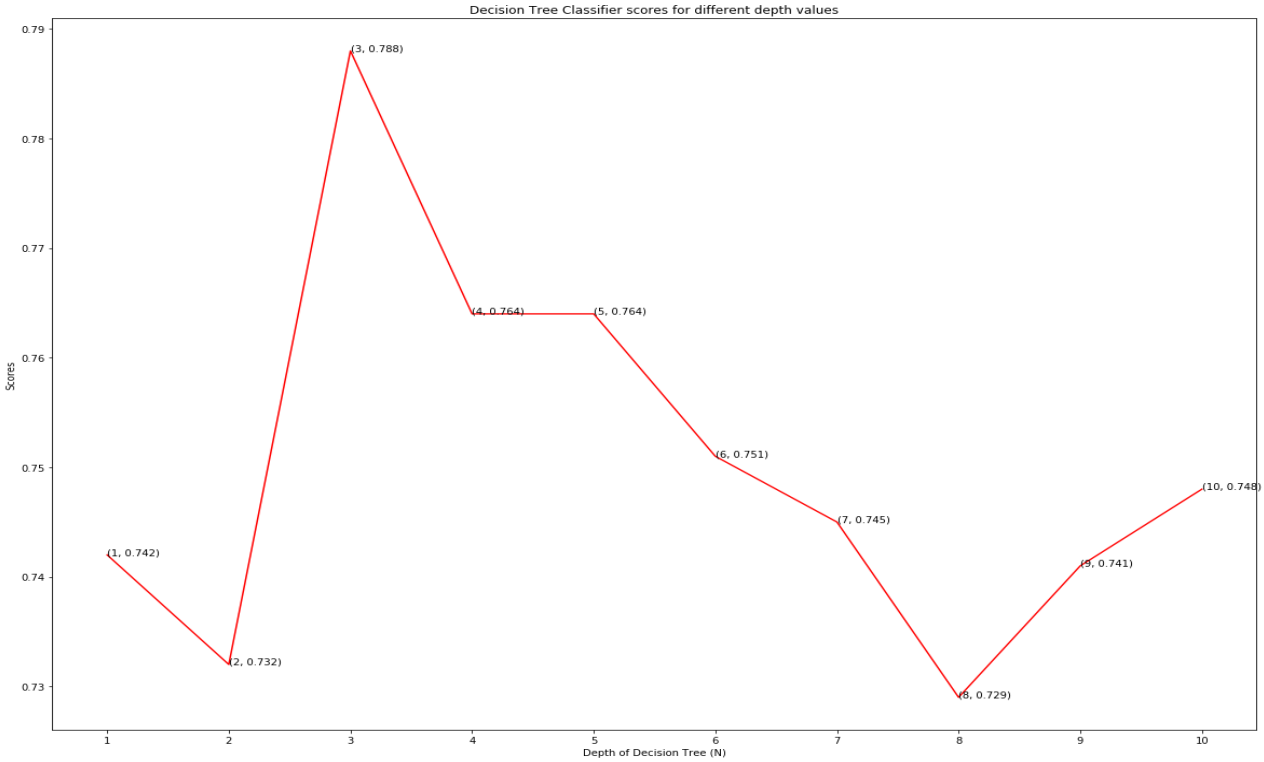
1. Correlation Matrix (Heat-Map)



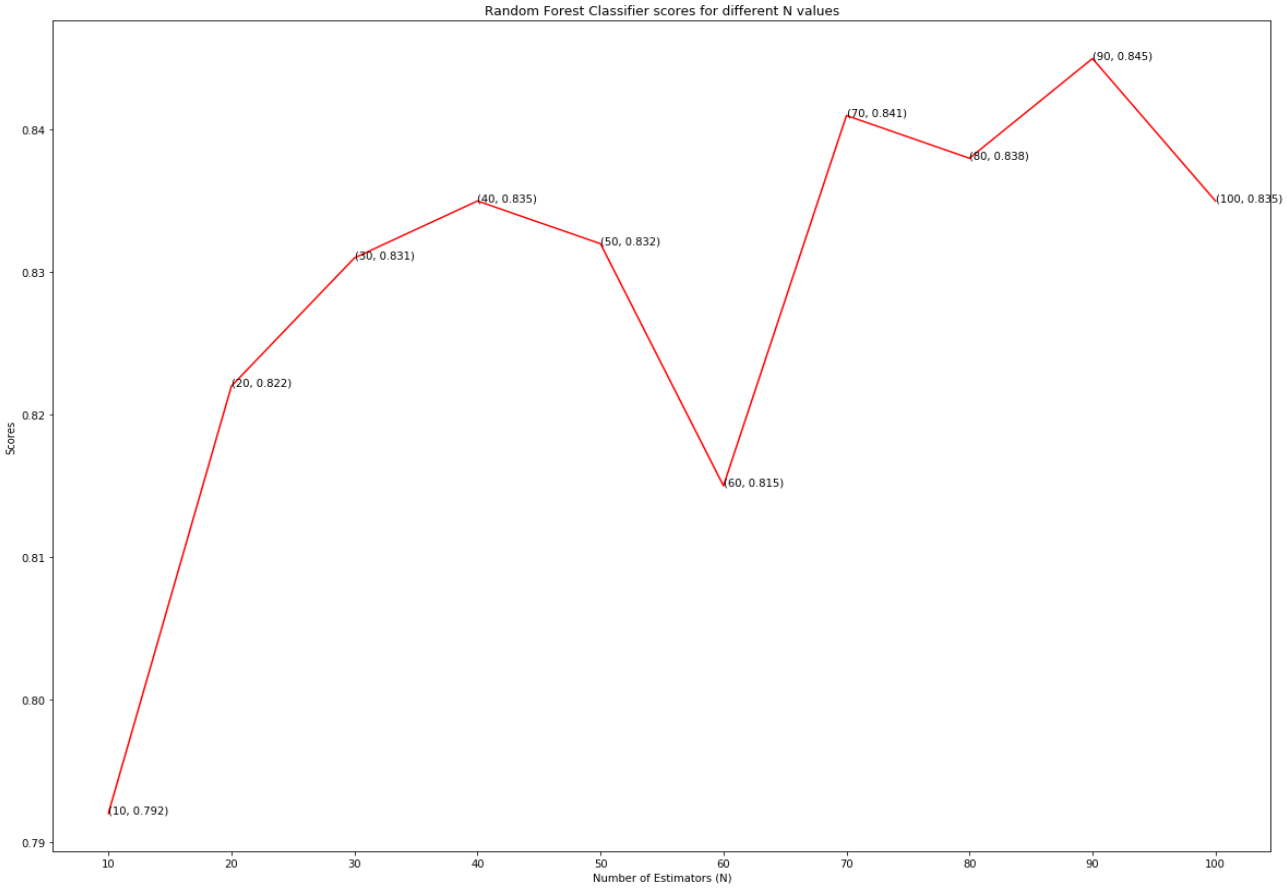
2. K Nearest Neighbour



3. Decision Tree



4. Random Forest



5. Future Scope

As illustrated before the system can be used as a clinical assistant for any clinicians. The disease prediction through the risk factors can be hosted online and hence any internet users can access the system through a web browser and understand the risk of heart disease. The proposed model can be implemented for any real time application .Using the proposed model other type of heart disease also can be determined. Different heart diseases asrheumatic heart disease, hypertensive heart disease, ischemic heart disease, cardiovascular disease and inflammatory heart disease can be identified. Other health care systems can be formulated using this proposed model in order to identify the diseases in the early stage. The proposed model requires an efficient processor with good memory configuration to implement it in real time. The proposed model has wide area of application like grid computing, cloud computing, robotic modelling, etc. To increase the performance of our classifier in future, we will work on ensembling two algorithms called Random Forest and Decision Tree.

6. Conclusion

This project, we introduce about the heart disease prediction system with different classifier techniques for the prediction of heart disease. The techniques are Random Forest and KNN ,have analyzed that the KNN has better accuracy as compared to Random Forest. Our purpose is to improve the performance of the Random Forest by removing unnecessary and irrelevant attributes from the dataset and only picking those that are most informative for the classification task.

We use some libraries provided by Python to implement this project. After the experiments, the algorithm of KNN Imputer gives us the best test accuracy, which is 84.48%. The reason why it outperforms others is that it is not limited to the property of the dataset. Though we get a good result of 84.48% accuracy, that is not enough because it cannot guarantee that no wrong diagnosis happens. To improve accuracy, we hope to require more dataset because 300 instances of dataset are not sufficient to do an excellent job. In the future, to predict disease we want to try different diseases such as lung cancer by using image detection. In this way, the dataset becomes complicated and we can apply convolutional neural network to make accuracy predictions.

7. **References**

[1]Soni, Jyoti, et al. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." International Journal of Computer Applications 17.8 (2011): 43-48.

[2] Dangare, Chaitrali S., and Sulabha S. Apte. "Improved study of heart disease prediction system using data mining classification techniques." International Journal of Computer Applications 47.10 (2012): 44-48.

[3] Uyar, Kaan, and Ahmet İlhan. "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks." Procedia computer science 120 (2017): 588-593.

[4] Kim, Jae Kwon, and Sanggil Kang. "Neural network-based coronary heart disease risk prediction using feature correlation analysis." Journal of healthcare engineering 2017 (2017).

[5] Baccouche, Asma, et al. "Ensemble Deep Learning Models for Heart Disease Classification: A Case Study from Mexico." Information 11.4

[6] https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[7] https://www.kaggle.com/ronitf/heart-disease-uci

[8] https://www.robots.ox.ac.uk/~az/lectures/ml/lect2.pdf

[9] https://nthu-datalab.github.io/ml/labs/03_Decision-Trees_RandomForest/03_Decision-Tree_Random-Forest.html

[10] https://www.kaggle.com/jprakashds/confusion-matrix-in-python-binaryclass

[11] scikit-learn, keras, pandas and matplotlib