Heart Disease Prediction Using Machine Learning Algorithm

Aman Singh
UG Student, Fisk University, Nashville, Tennessee, 37208

Abstract

Heart disease (HD) is the major cause of the human death. Many people die due to this disease. Many researchers have been discovering new technologies to prognosticate the disease early before it's too late for helping healthcare as well as people. These processes are still under the research phase. Machine Learning (ML) is faster-emerging technologyof Artificial Intelligence (AI) that contributes various algorithms for HD. Based on the problem that has been proposed, Machine Learning provides different classification algorithms to divine the probability of patient having HD. For predicting HD, a lot of research scholars contribute their effort in this work using various techniques and algorithms such as Decision Tree (DT), KNN (K- Nearest Neighbor), etc. To give some effort to this work, we are going to implement three ML algorithms that are K neighbor Classifier, Random Classifier, and Decision Tree Classifier. We are using the UCI repository HD dataset to train a model by comparing DT, KNN, and Random Forest Classifier algorithms for our final output. The dataset contains 303 instances with 14 attributes that help to train a prediction model that will be deployed for prediction. This project aims to build and deploy an efficient prediction model. An HDP Model is built by using a Random Classifier algorithm that provides 83.51% accuracy among others. Each phase is efficiently done. With the help of requirement analysis and project plan, system design, database design, testing plan, identifying features and functionalities, and system validation and deployment, this project is created. The limitation in this project is to only predict the presence of heart disease, not identify the type of HD. In the future, we can enhance the project by appending more detailed predictions of HD of patients and incorporating smart wearable devices that integrate into Hospital Emergency systems.

Keywords: Machine Learning (ML), Decision Tree (DT), Random Forest Classifier, K-Nearest Neighbor, HeartDisease, Classification.

1. INTRODUCTION

ML is an emerging application of AI that uses different analytics and statistical techniques to improve the performance of machine learning from old data. It enables a particular machine to learn from a database and enhance the performanceby experience. It helps to build an intelligent machine to solve a specific problem. ML solved various complex problems that don't solve by statistical algorithms. ML provides dynamic algorithms which are being without an explicit program to build an intelligent machine that can ease various difficult problems. ML solved the different types ofproblems which are categorized into three parts. Supervised, unsupervised, and reinforcement type problems.In supervised, there are two types of problems such as.section and regression problems. In unsupervised, there is a clustering type problem that can solve ML algorithms. ML assigned different algorithms based on their type ofproblem. ML project is done by the following steps:

- Defines a problem statement.
- Classifying the problem into ML problems.

- Selecting suitable ML algorithms based on their type of problems.
- Collecting and cleaning the data.
- Training a Model from data.
- Test the Model from test data
- Evaluate a model from their accuracy.

This work is closely related to the supervised problem of ML. However, many researchers have also solved this problem from unsupervised and NN. NN is a subset of ML that solved the complex problem which does not solve by normal ML. this project willsolve by three classification algorithms of supervised ML such as DT and Random Forest and K-NN. This system will perform Python programming language which is handled entire development of this system using its ML's libraries. At the initial phase, a Heart Disease Prediction Model (HDPM) will be built by one of three mentioned algorithms through the comparison of them. This process is done through the ML Process to build a model.

1.1 Background to the project

According to WHO, HDs have become the leading cause of death in the last fifteenyears (WHO, 2019). Millions of humans are having a HD every year so that the HD takes placed the biggest killer of people in the world. According to analyzation of WHO, twelve million people are death due to HD in worldwide. One person dies almost every 34 seconds from HD. Diagnosis of HDs is an essential task and yet intricate task to perform accurately and efficiently in the hospital and clinic.This proposed system can reserve problems by accurately predicting the presence of HD in the patient. ML is an emerging technology of AI that solved various types of classification problems by producing accurate output. ML algorithms are applied to forecast the HD of the patient. This system is useful to support the decision-making of doctors and healthcare members inhospitals.

1.2 Problem context

HD defines a condition that affects a heart. HD contains differences diseases such as coronary artery disease (CAD), Congenital HD, Mitral Value Prolapse, Arrhythmia, Pulmonary Stenosis, Dilated Cardiomyopathy, Heart Failure, Hypertrophic Cardiomyopathy, and Myocardial Infarction. One of them, cardiovascular disease (CVD) is one of the main diseases of the heart that refers to the condition of obstructed blood vessels that can be happened a stroke and heart attack. Another form of HD can be rhythm, heart's muscle, etc. (Mayo Clinic, 2019)

CVDs are one of the major cause of people death globally. Many people have died from CVDs compare to other cause. In 2016, due to CVDs, an estimated 17.9 million human died. It's illustrating 31% of human deaths all over the world. Stroke and heart attack have occupied 85% of these deaths. (World Health Organization, 2019)

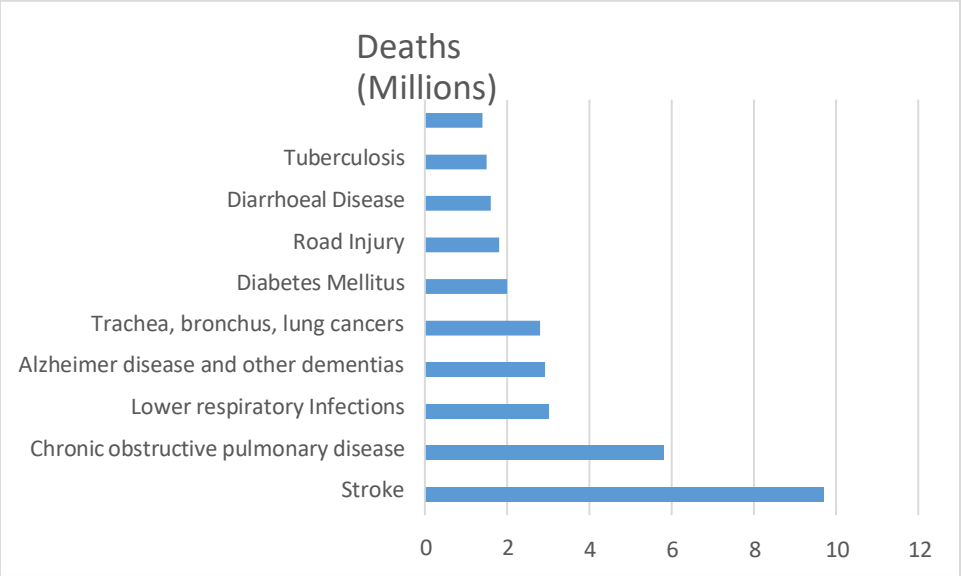


Figure 1: Top 10 Global causes of deaths (WHO, 2019)

In Figure-1.1, the WHO analysis the data about causes of deaths in between 2000-2016 and result clearly shows that the causes of heart disease's death is higher than other causes of death.

In 2017, the latest fact data of Word Health Organization (WHO) published that Nepal has reached 18.72% or 30,559 deaths from Coronary HD. The rate of age fixed death is 158.35 out of 100,000 population and world rank is #41. (World Life Expectancy, 2019)

According to The Heart Foundation, 13% of men and 10% of women are died due to HD in Australia. In 2017, Whilst HD had 18,590 deaths. So that HD was a one four death of cause factor in 2017. (The Heart Foundation., 2019)

So, Nepal government also needs to use this system to aware the patient before being critical situation. This system provides accurate result that help to less worry about the doctor's negligence.

2. SYSTEM ARCHITECTURE

2.1 Introduction

The system architecture is like a blueprint of any object. It is a conceptual model to integrate between business logic and physical system in an organized way. It demonstratesthe structure, view, behavior, features, and functionalities of the system. It is the way of portraying the desired system in visualizing a way to well understand f

or people. The System architecture is the foundational orchestrate of a system that incorporated in its elements, their relationships of elements, and the science of its design (MITRE, 2019).

Here, in the Heart Disease Prediction Model, First, we are visualizing our data with different features and providing a correlation matrix with the provided features in our dataset. After that, we will be implementing different Machine learning models, like K-Nearest Neighbor, Random Forest Classifier and Decision Tree Classifier. After implementing those models, we will be retrieving our best possible score to predict the probability if a person has a heart disease or not.

2.2 Abstract Architecture

Abstract Architecture of HDPS has demonstrated the comprehensive structure of the system that easily understands all components and their relationship. The way to represent the system as a structural way can be beneficial to understand the system for technical and non-technical users. The abstract architecture consists of a system, database, and interface design to visualize the better way of viewing. It describes the overall design of the system to easily understand the system elements, feature and functionalities, and behavior of the system.

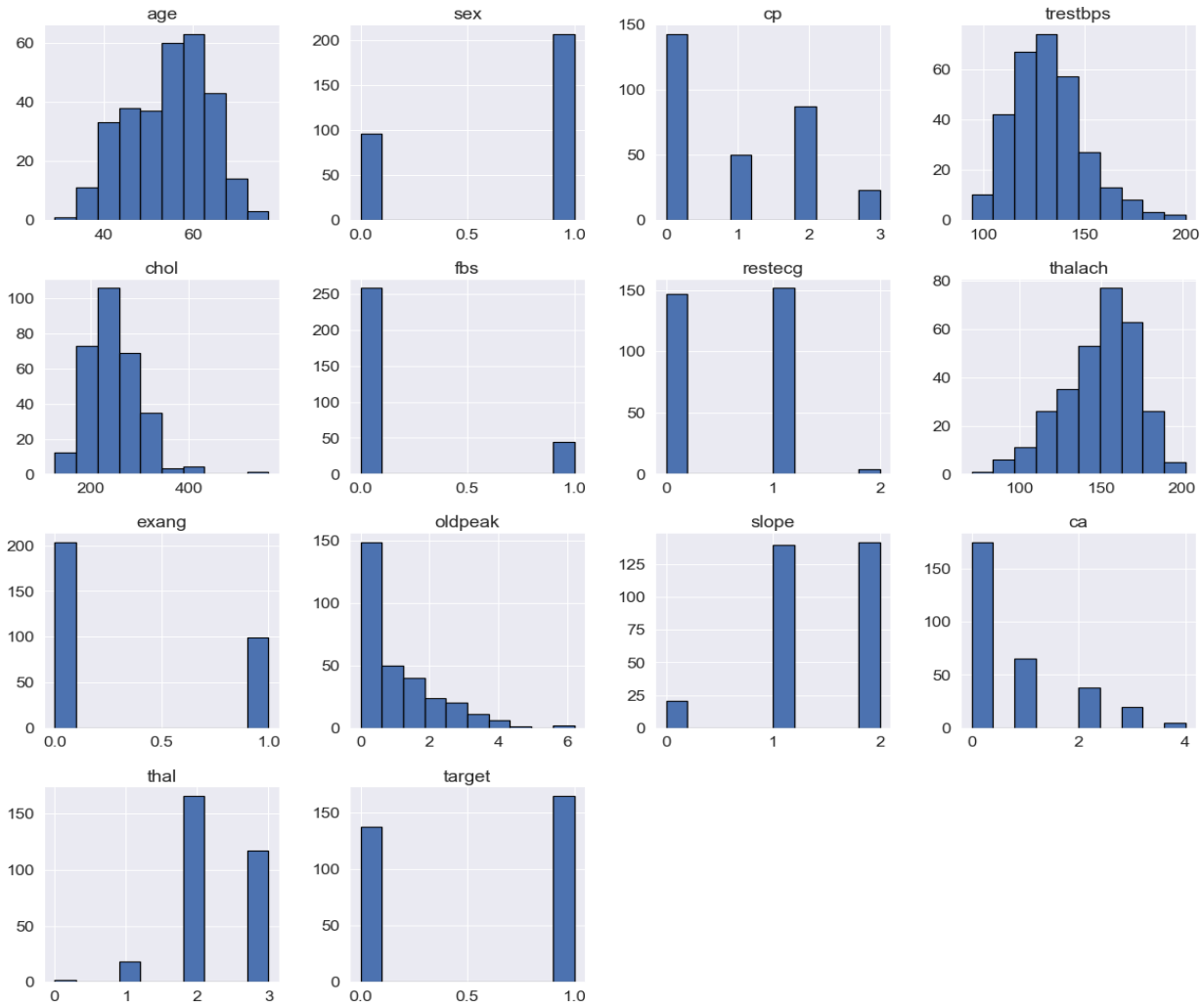
3. The Design Methodology:

Model is trained to predict whether a person has heart disease or not based on the following common features as input:

- age
- gender
- chest pain
- blood pressure
- cholesterol level
- max heart rate

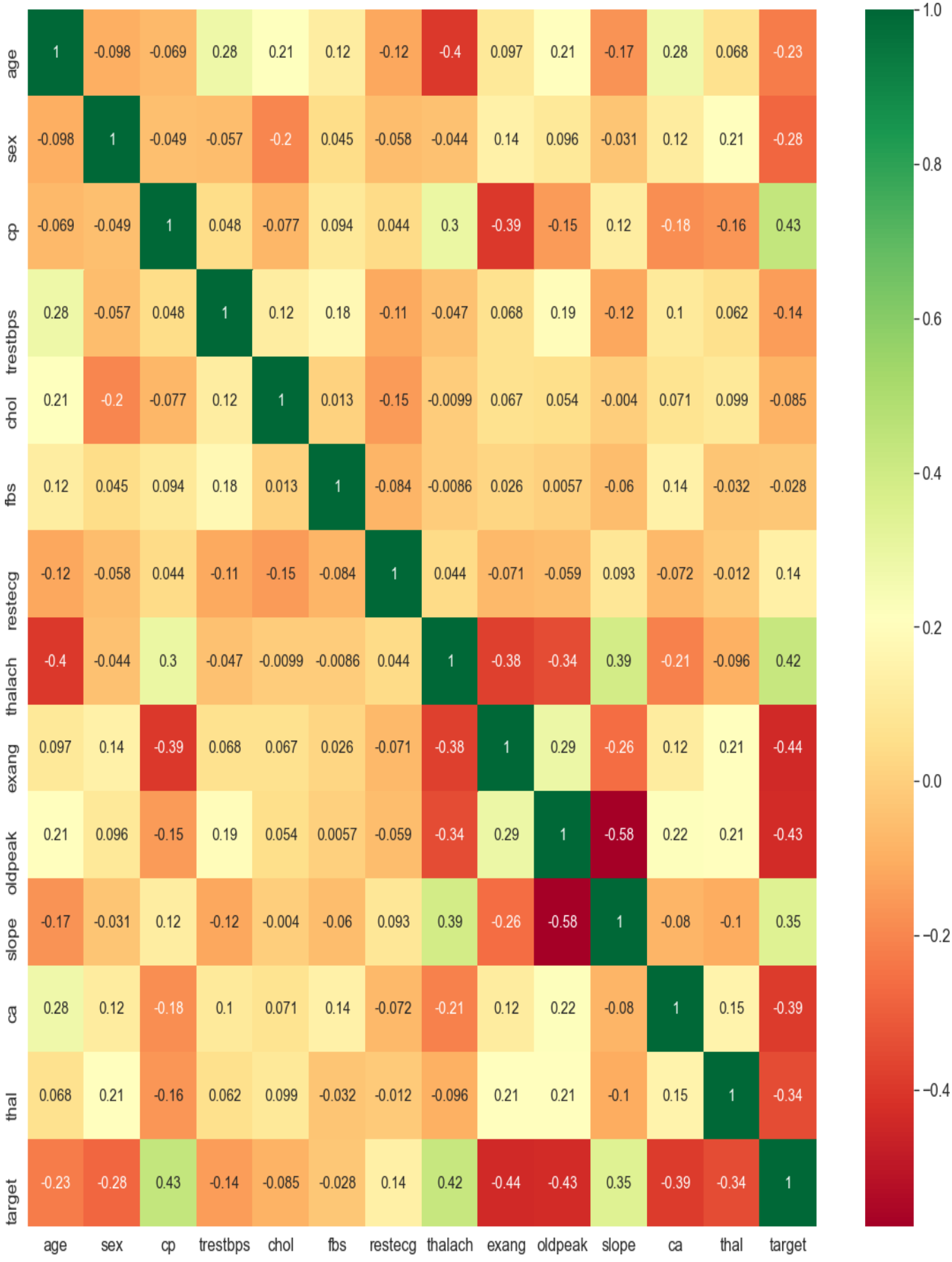
3.1 Visualizing our dataset.

Now let us see various visual representations of the data to understand more about relationship between various features.



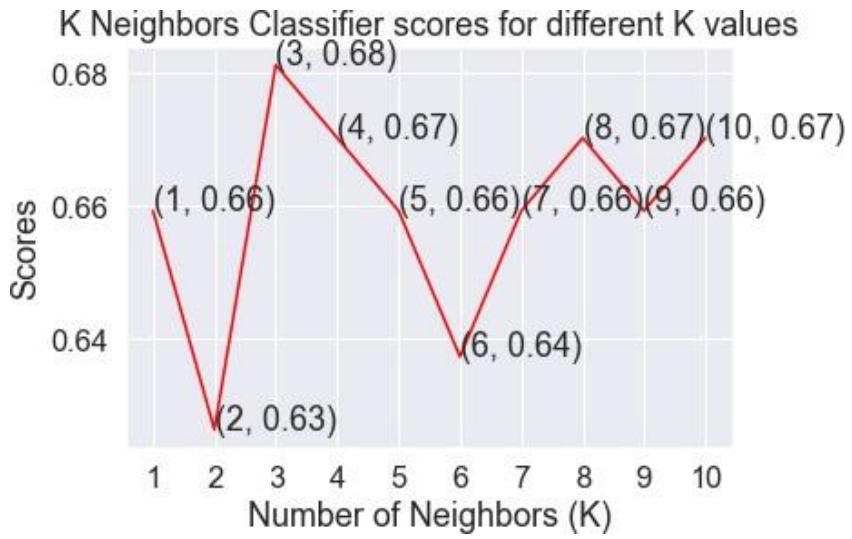
3.2 Correlation Matrix

One of the best ways to compare relationship between various features is to look at the correlation matrix between those features.


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
age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal target
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

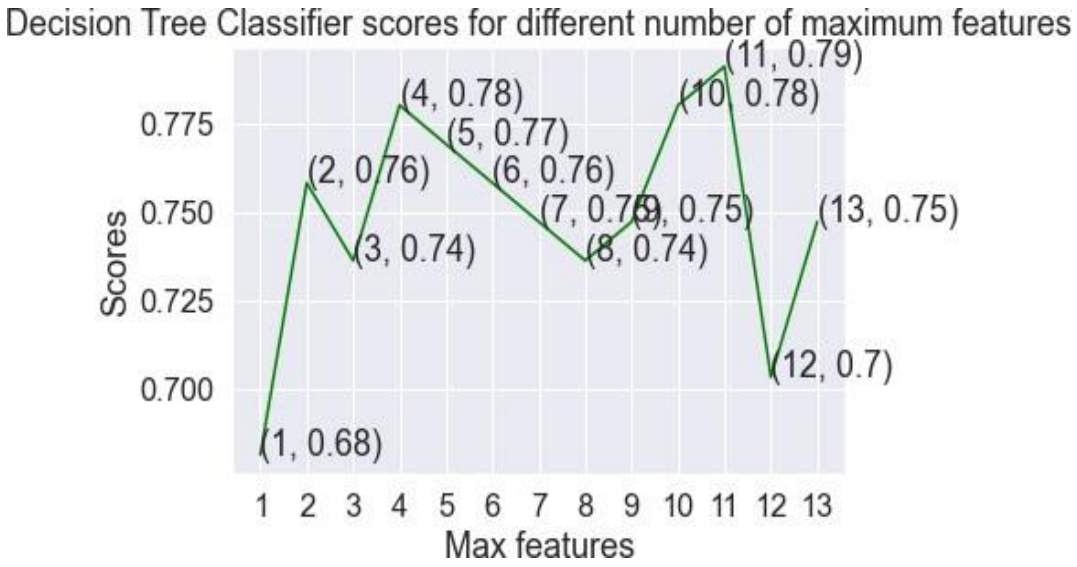
3.3 Implementing Machine Learning Models

1. K-Nearest Neighbor Classifier



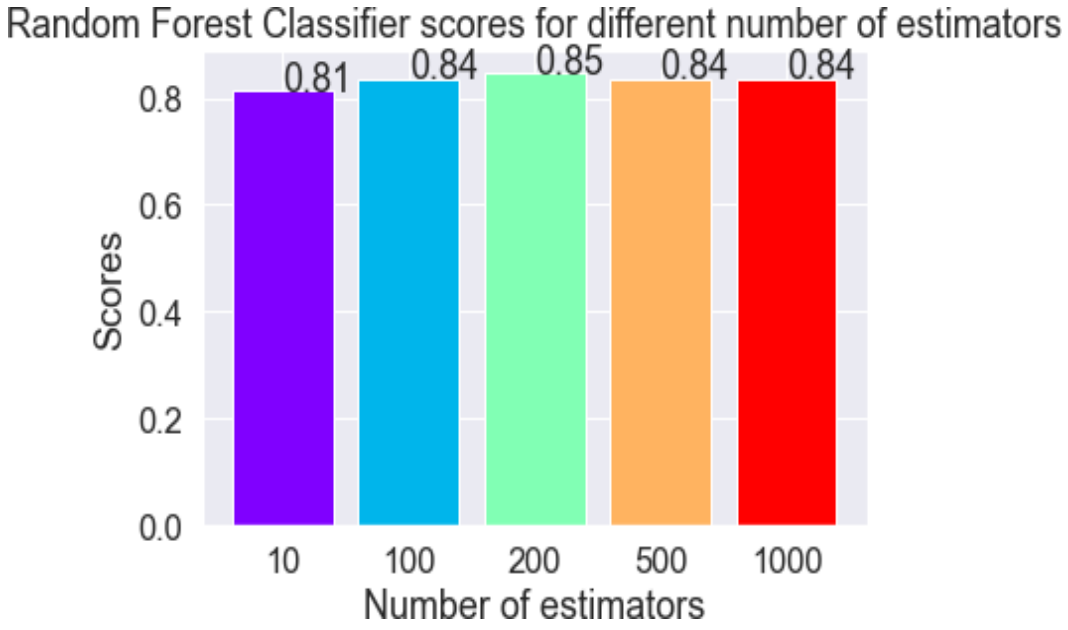
The score for K Neighbors Classifier is 60.43956043956044% with 8 neighbors.

2. Decision Tree Classifier



The score for Decision Tree Classifier is 70.32967032967034% with [2, 4, 11] maximum features

3. Random Forest Classifier



The score for Random Forest Classifier is 83.51648351648352% with [100, 200] estimators.

4. CONCLUSION AND REFLECTIONS

In this project, I used Machine Learning to find or predict whether a person is suffering from a heart disease or not. After importing the data, I analyzed it using plots. Then, I did generate dummy variables for categorical features and scaled other features. After that, I applied 3 Machine Learning algorithms, K Neighbors Classifier, Decision Tree Classifier and Random Forest Classifier. I varied parameters across each model to improve their scores. At the end, Random Classifier achieved the highest score of 83.5%.

5. Future Work

Future enhance of the Heart Disease Prediction Model is to predict a specific HD type such Heart attracts, CVD, CAD, etc. the potential of the HDPS in a different area are hospital, Clinic, smartphone, smart wear, hospital/police emergency system and integrate with fitness mobile application. We will integrate this model in hospital and clinic system to predict heart disease. We will implement this HDP Model into smart wears to detect essential attributes of HD and suggest to the precaution of HD. we will also apply this model into a mobile app to easily test ourselves HD. we will integrate smart wear to the hospital and police emergency system to save the life of the patient at the emergency condition.

6. REFERENCES

Segue Technologies, 2015. *The Benefits of Adhering to Software DevelopmentMethodology Concepts.* [Online]
Available at: https://www.seguetech.com/benefits-adhering-software-development-methodology-concepts/
[Accessed 28 01 2019].

Abhishek, t., 2013. Heart Disease Prediction System Using Data Mining Techniques. *Oriental Scientific Publishing Co., India,* 6(4), pp. 457-466.

Afroz, R. R. a. F., 2013. Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis. *Journal of Software Engineeringand Applications,* 6(3), pp. 85-97.

Aparna K., D. R. N. C. S. P. I. S. S. D. K. V., 2014. Disease Prediction in Data Mining Techniques. *InternatIonal Journal of Computer SCIenCe and teChnology,* 5(2), pp. 246-249

Arsenault, R., 2019. *Sublime Text.* [Online]
Available at:
https://www.utest.com/tools/sublime-text
[Accessed 30 01 2019].

Asthana, S., 2018. *Essential libraries for Machine Learning in Python.* [Online] Available at: https://medium.freecodecamp.org/essential-libraries-for-machine-learning-in-python-82a9ada57aeb
[Accessed 08 02 2019].

Bahrami, B. a. S. M., 2015. Prediction and Diagnosis of Heart Disease by Data Mining Techniques. *Journal of Multidisciplinary Engineering Science and Technology (JMEST),*2(2), pp. 164-168.

Chala Beyene, P. K., 2018. "Survey on Prediction and Analysis the Occurrence of HeartDisease Using Data Mining Techniques. *International Journal of Pure and Applied Mathematics,* 118(8), pp. 165-174.

Dangare Chaitrali S., P. A. S. S., 2012. Improved Study of Heart Disease Prediction System using Data Mining Classification Technique. *International Journal of ComputerApplications (0975 – 888) ,* 47(10), pp. 44-48.