HEARTH DISEASE PREDICTOR

MAJOR PROJECT REPORT

Submitted by
VINAYAK SHARMA
2016-333-061

in partial fulfillment for the award of the degree of
BACHELOR OF TECHNOLOGY IN ELECTRONICS AND COMMUNICATION ENGINEERING

Under the supervision of
Mr. NASEEM RAO

Department of Computer Science
JAMIA HAMDARD
(Hamdard University)
New Delhi-110062
(2020)

DECLARATION

I, **Mr. Vinayak Sharma** a student of **Bachelor of Technology (Electronics & Communication Engineering), Enrolment No: 2016-333-061** hereby declare that the dissertation entitled **"Heart Disease Predictor"** which is being submitted by me to the Department of Computer Science, Jamia Hamdard, New Delhi in partial fulfillment of the requirement for the award of the degree of **Bachelor of Technology (Electronics & Communication Engineering)** is my original work and has not been submitted anywhere else for the award of any Degree, Diploma, Associate ship, Fellowship or other similar title or recognition.

Date:

Place: Jamia Hamdard, New Delhi **(Signature and Name of the Applicant)**

ACKNOWLEDGEMENT

I, Vinayak Sharma feels pleasure to bring out this project named “Heart Disease predictor”. I take this opportunity to thanks everyone who guided me through and many people who knowingly and willingly helped me, to complete my project. First of all, let us thank God for all the blessings, which carried us through all these years. I extend my utmost gratitude to **Mr. NASEEM RAO**, my project supervisor who always stood by my side and guided, appreciated and encouraged me to get into more and more ventures. Continuing the same, he enlightened me in the various stages during the development of this project and provided me with many insights and useful examples, which proved to be of immense help in successful completion of this project.

I extend my sincere gratitude to my teachers and guide who made unforgettable contribution. I thank all the non-teaching staff of our institution that was always ready to help in whatever way they could.

Vinayak sharma

Place: JAMIA HAMDARD, NEW DELHI

TABLE OF CONTENT

HEART
DISEASE
PREDICTOR

Background

Among all fatal disease, heart attacks diseases are considered as the most prevalent. Medical practitioners conduct different surveys on heart diseases and gather information of heart patients, their symptoms and disease progression. Increasingly are reported about patients with common diseases who have typical symptoms. In this fast moving world people want to live a very luxurious life so they work like a machine in order to earn lot of money and live a comfortable life therefore in this race they forget to take care of themselves, because of this there food habits change their entire lifestyle change, in this type of lifestyle they are more tensed they have blood pressure, sugar at a very young age and they don't give enough rest for themselves and eat what they get and they even don't bother about the quality of the food if sick the go for their own medication as a result of all these small negligence it leads to a major threat that is the heart disease.

The term 'heart disease' includes the diverse diseases that affect heart. The number of people suffering from heart disease is on the rise (health topics, 2010). The report from world health organization shows us a large number of people that die every year due to the heart disease all over the world. Heart disease is also stated as one of the greatest killers in Africa.

Data mining has been used in a variety of applications such as marketing, customer relationship management, engineering, and medicine analysis, expert prediction, web mining and mobile computing. Of late, data mining has been applied successfully in healthcare fraud and detecting abuse cases.

OBJECTIVE

Main Objective

The main objective of this research is to develop a heart prediction system. The system can discover and extract hidden knowledge associated with diseases from a historical heart data set

Heart disease prediction system aims to exploit data mining techniques on medical data set to assist in the prediction of the heart diseases.

Specific Objective

- Provides new approach to concealed patterns in the data.
- Helps avoid human biasness.
- To implement Naïve Bayes Classifier that classifies the disease as per the input of the user.
- Reduce the cost of medical tests.

Justification

Clinical decisions are often made based on doctor's insight and experience rather than on the knowledge rich data hidden in the dataset. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. The proposed system will integrate clinical decision support with computer-based patient records (Data Sets). This will reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge rich environment which can help to significantly improve the quality of clinical decisions.

Scope and Limitation

Scope

Here the scope of the project is that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions

Limitations

Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. The automation of the same would be highly beneficial. Clinical decisions are often made based on doctor's intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Data mining have the potential to generate a knowledge-rich environment which can help to significantly
improve the quality of clinical decisions.

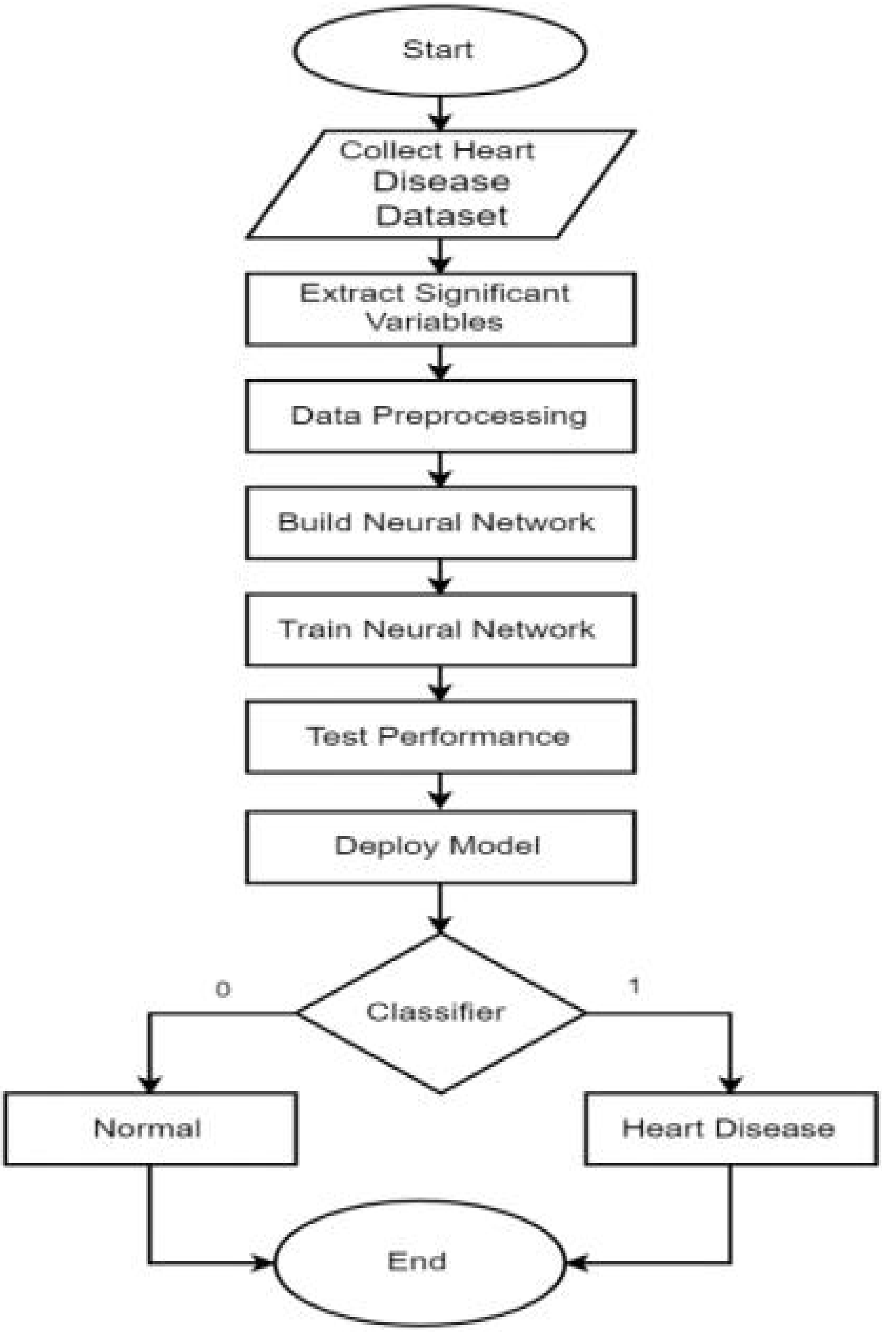
INTRODUCTION

Numerous studies have been done that have focus on diagnosis of heart disease. They have applied different data mining techniques for diagnosis & achieved different probabilities for different methods. (Polaraju, Durga Prasad, & Tech Scholar, 2017) proposed Prediction of Heart Disease using Multiple Regression Model and it proves that Multiple Linear Regression is appropriate for predicting heart disease chance. The work is performed using training data set consists of 3000 instances with 13 different attributes which has mentioned earlier. The data set is divided into two parts that is 70% of the data are used for training and 30% used for testing. (Deepika & Seema, 2017) focuses on techniques that can predict chronic disease by mining the data containing in historical health records using Naïve Bayes, Decision tree, Support Vector Machine (SVM) and Artificial Neural Network (ANN). A comparative study is performed on classifiers to measure the better performance on an accurate rate. From this experiment, SVM gives highest accuracy rate, whereas for diabetes Naïve Bayes gives the highest accuracy. (Beyene & Kamat,)recommended different algorithms like Naive Bayes, Classification Tree, KNN, Logistic Regression, SVM and ANN. The Logistic Regression gives better accuracy compared to other algorithms. (Beyene & Kamat, 2018) suggested Heart Disease Prediction System using Data Mining Techniques. WEKA software used for automatic diagnosis of disease and to give qualities of services in healthcare centers. The paper used various algorithms like SVM, Naïve Bayes, Association rule, KNN, ANN, and Decision Tree. The paper recommended SVM is effective and provides more accuracy as compared with other data mining algorithms. Chala Beyene recommended Prediction and Analysis the occurrence of Heart Disease Using Data Mining Techniques. The main objective is to predict the occurrence of heart disease for early automatic diagnosis of the disease within result in short time. The proposed methodology is also critical in healthcare organization with experts that have no more knowledge and skill. It uses different medical attributes such as blood sugar and heart rate, age, sex are some of the attributes are included to identify if the person has heart disease or not. Analyses of data set are computed using WEKA software. (Soni, Ansari, & Sharma, 2011) proposed to use non- linear classification algorithm for heart disease prediction. It is proposed to use bigdata tools such as Hadoop Distributed File System (HDFS), Map reduce along with SVM for prediction of heart disease with optimized attribute set. This work made an investigation on the use of different data mining techniques for predicting heart diseases. It suggests to use HDFS for storing large data in different nodes and executing the prediction algorithm using SVM in more than one node simultaneously using SVM. SVM is used in parallel fashion which yielded better computation time than sequential SVM. (Science & Faculty, 2009) suggested heart disease prediction using data mining and machine learning algorithm. The goal of this study is to extract hidden patterns by applying data mining techniques. The best algorithm J48 based on UCI data has the highest accuracy rate compared to LMT. (Purushottam, Saxena, & Sharma, 2016) proposed an efficient heart disease prediction system using data mining. This system helps medical practitioner to make effective decision making based on the certain parameter. By testing and training phase a certain parameter, it provides 86.3% accuracy in testing phase and 87.3% in training phase. This paper proposed data mining techniques to predict the disease. It is intended to provide the survey of current techniques to extract information from dataset and it will useful for healthcare

practitioners. The performance can be obtained based on the time taken to build the decision tree

WORKING

This will be the proposed flow chart that the system will look like

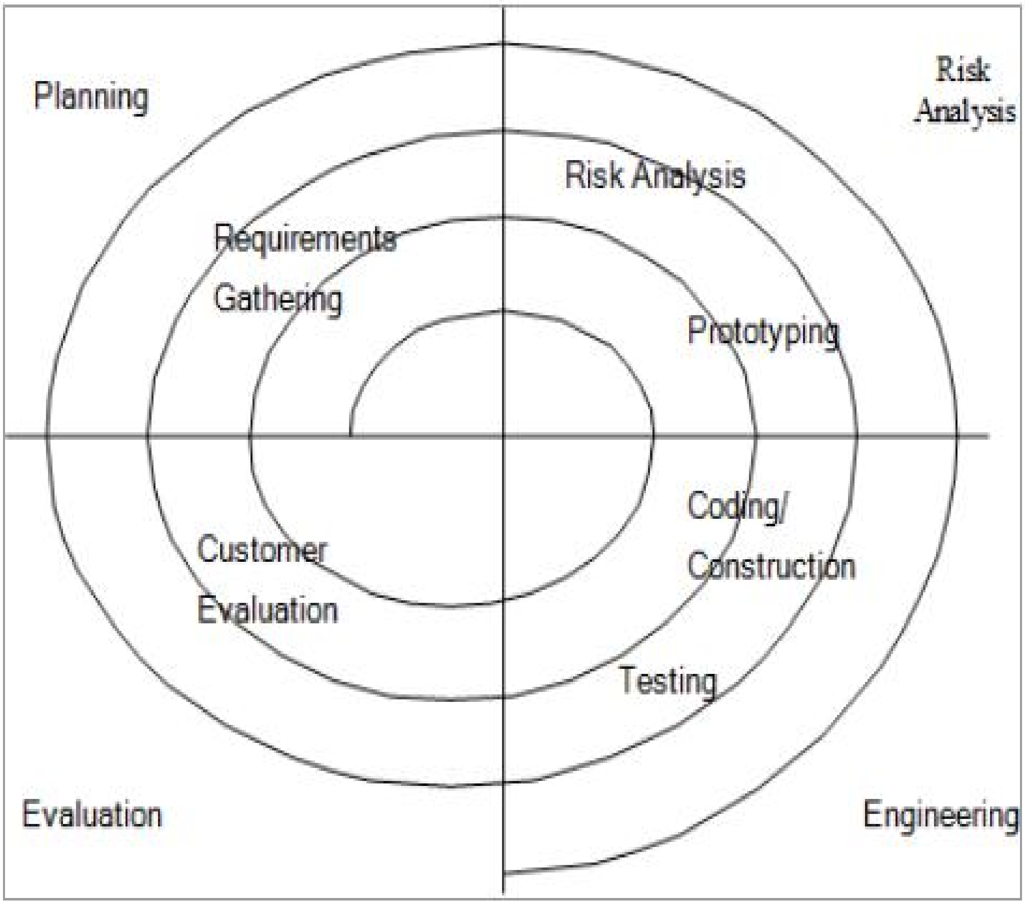


Research Design

I will be using the experimental type of research design. It is a quantitative research method. Basically, it is a research conducted with a scientific approach, where a set of variables are kept constant while other set of variables are being measured as the subject of the experiment. This is more practically while conducting face recognition and detection as it monitors the behaviours and patterns of a subject to be used to acknowledge whether the subject matches all details presented and cross checked with previous data. It is an effect research method as it is time bound and focuses on the relationship between the variables that give actual results.

System Development Methodology

The methodology of software development is the method in managing project development. There are many models of the methodology are available such as Waterfall model model, Incremental model, RAD model, Agile model, Iterative model and Spiral model. However, it still need to be considered by developer to decide which is will be used in the project. The methodology model is useful to manage the project efficiently and able to help developer from getting any problem during time of development. Also, it help to achieve the objective and scope of the projects. In order to build the project, it need to understand the stakeholder requirements. Methodology provides a framework for undertaking the proposed DM modeling. The methodology is a system comprising steps that transform raw data into recognized data patterns to extract knowledge for
users.



There are four phases that involve in the spiral model:

1) Planning phase

Phase where the requirement are collected and risk is assessed. This phase where the title of the project has been discussed with project supervisor. From that discussion, Heart Prediction System has been proposed. The requirement and risk was assessed after doing study on existing system and do literature review about another existing research.

2) Risk analysis Phase

Phase where the risk and alternative solution are identified. A prototype are created at the end this phase. If there is any risk during this phase, there will be suggestion about alternate solution.

3) Engineering phase

At this phase, a software are created and testing are done at the end this phase.

4) Evaluation phase

At this phase, the user do evaluation toward the software. It will be done after the system are presented and the user do test whether the system meet with their expectation and requirement or not. If there is any error, user can tell the problem about system.

ALGORITHMS

Naïve Bayesian

It is a probabilistic classifier based on Bayes' theorem specified by the prior probabilities of its root nodes. The Bayes theorem is given in Equation 1 and normalization constant is given in Equation 2. It proves to be an optimal algorithm in terms of minimization of generalized error. It can handle statistical based machine learning for feature vectors and assign the label for feature vector based on maximal probable among available classes {XX1, X2..., XM}. It means that feature "y" belongs to Xiclass, when posterior probability is maximum ie Max. The Bayesian classification problem may be formulated by a posterior probabilities that assign the class label ωi to sample X such that is maximal. The Bayesian classification problem may be formulated by a-posterior probabilities that assign the class label ωi to sample X such that is maximal.

Likelihood

Prior

$$P(X_i|\underline{y}) = \frac{p(\underline{y}|X_i)P(X_i)}{p(\underline{y})}$$

Normalization Constant

$$p(\underline{y}) = \sum_{i=1}^{2} p(\underline{y}|X_i)P(X_i)$$

$P(X1, X2..., Xn | Y) = P(X1 | Y) P(X2 | Y) ... P(Xn | Y)$

Decision Trees.

The decision tree approach is more powerful for classification problems. There are two steps in this technique building a tree & applying the tree to the dataset. There are many popular decision tree algorithms CART, ID3, C4.5, CHAID, and J48. From these J48 algorithm is used for this system. J48 algorithm uses pruning method to build a tree. Pruning is a technique that reduces size of tree by removing over fitting data, which leads to poor accuracy in predications. The J48 algorithm recursively classifies data until it has been categorized as perfectly as possible. This technique gives maximum accuracy on training data. The overall concept is to build a tree that provides balance of flexibility & accuracy.

Random Forest

Random forests is a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is. Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a pretty good indicator of the feature importance.Random forests has a variety of applications, such as recommendation engines, image classification and feature selection. It can be used to classify loyal loan applicants, identify fraudulent activity and predict diseases. It lies at the base of the Boruta algorithm, which selects important features in a dataset.

Support Vector Machine

“Support Vector Machine”(SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well

K-Nearest Neighbour (KNN) Classification

When KNN is used for classification, the output can be calculated as the class with the highest frequency from the K-most similar instances. Each instance in essence votes for their class and the class with the most votes is taken as the prediction.

It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data (as opposed to other algorithms such as GMM, which assume a Gaussian distribution of the given data).We are given some prior data (also called training data), which classifies coordinates into groups identified by an attribute.

HARDWARE AND SOFTWARE REQUIREMENT

Introduction:

The following subsections of the Hardware and SRS document provide an overview:

Hardware

- RAM: 2 GB or more
- Processor: Intel Core I3 or higher power
- 512 KB Cache
- Keyboard: Normal or Multimedia
- Mouse: Compatible mouse

Software

- Python 3.6

Python 3.6

Python is a general-purpose programming language. Hence, you can use this language for developing mobile, pc and web applications. Also, Python can be used for developing complex scientific and numeric applications. Python is designed with features which can be facilitate data analysis and visualization. In order to make these applications there are various versions of Python Programming Language software are available on the internet. For this project, I have used Python 3.6

IMPLEMENTATION

This Python 3 environment comes with many helpful analytics libraries installed.It is defined by the kaggle/python docker image: https://github.com/kaggle/docker-python

Library used:

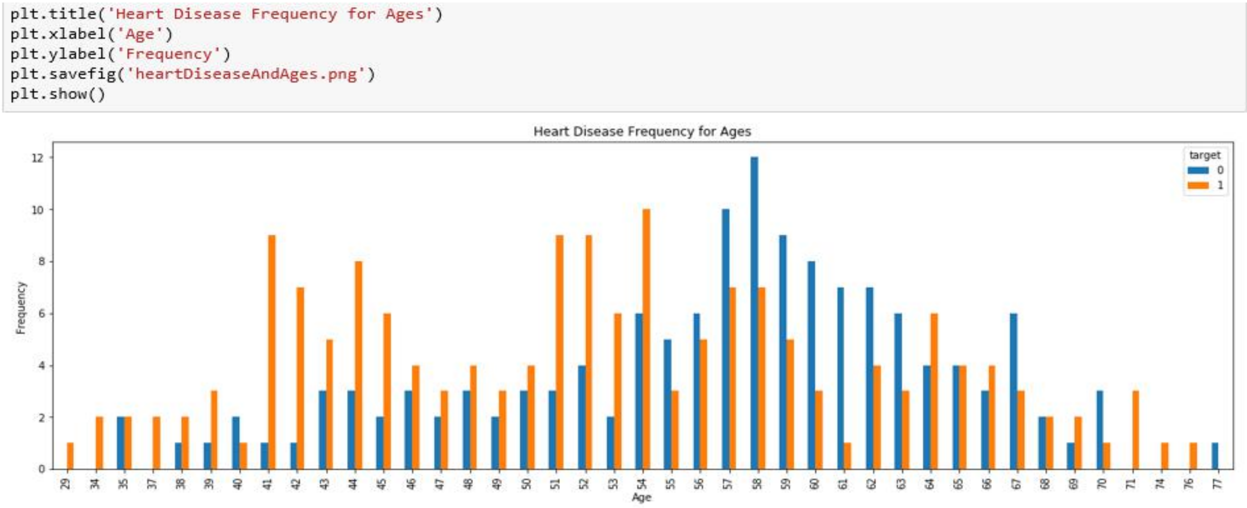
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
```

Data contains

- age - age in years
- sex - (1 = male; 0 = female)
- cp - chest pain type
- trestbps - resting blood pressure (in mm Hg on admission to the hospital)
- chol - serum cholestoral in mg/dl
- fbs - (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- restecg - resting electrocardiographic results
- thalach - maximum heart rate achieved
- exang - exercise induced angina (1 = yes; 0 = no)
- oldpeak - ST depression induced by exercise relative to rest
- slope - the slope of the peak exercise ST segment
- ca - number of major vessels (0-3) colored by flourosopy
- thal - 3 = normal; 6 = fixed defect; 7 = reversable defect
- target - have disease or not (1=yes, 0=no)

Screenshots

Heart disease frequency for ages

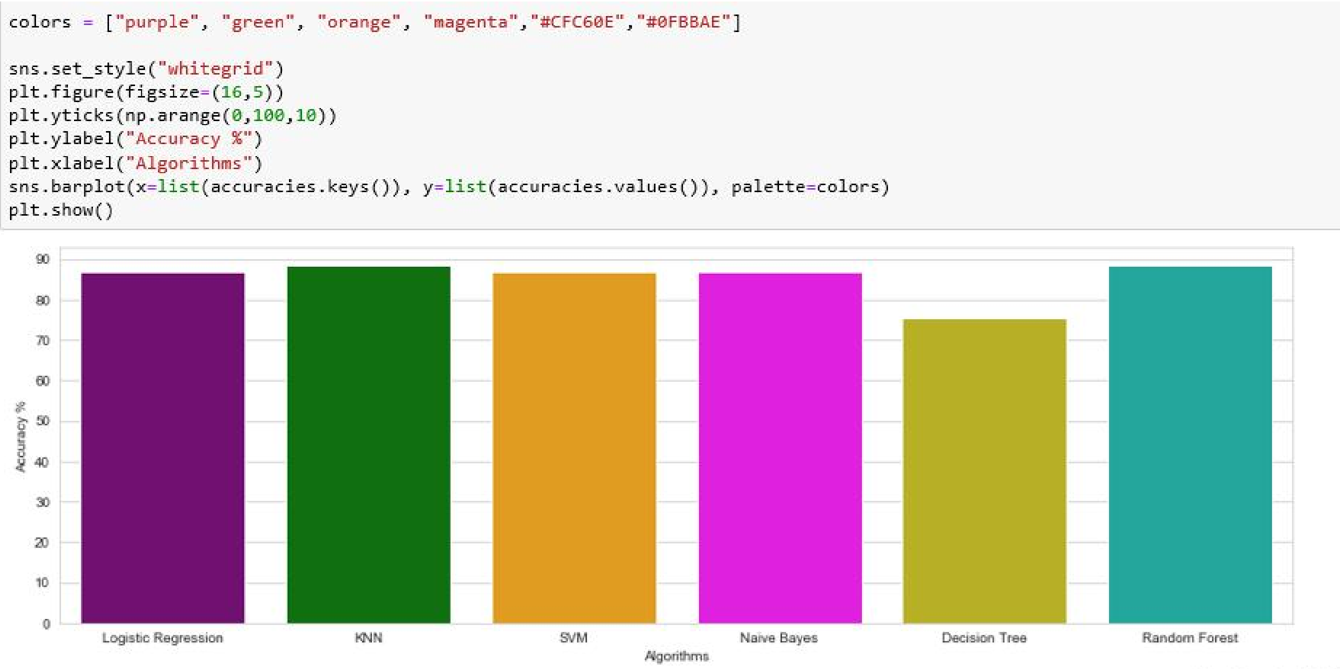


```
plt.title('Heart Disease Frequency for Ages')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.savefig('heartDiseaseAndAges.png')
plt.show()
```

Results from different algorithms

Our models work fine but best of them are KNN and Random Forest with 88.52% of accuracy. Let's look their confusion matrices

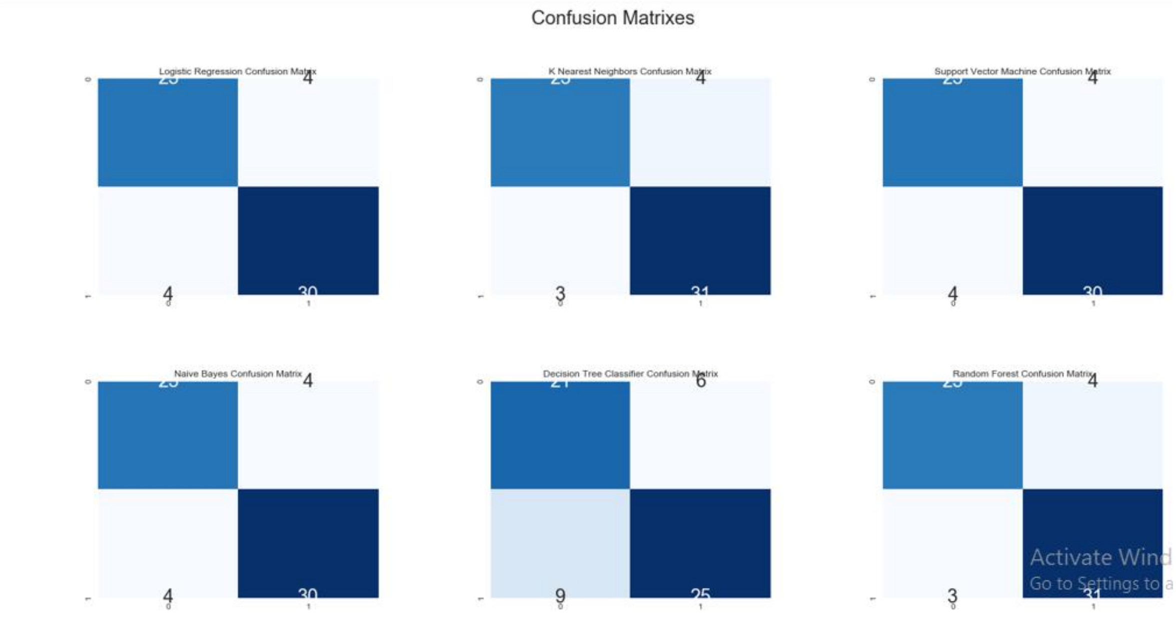
.



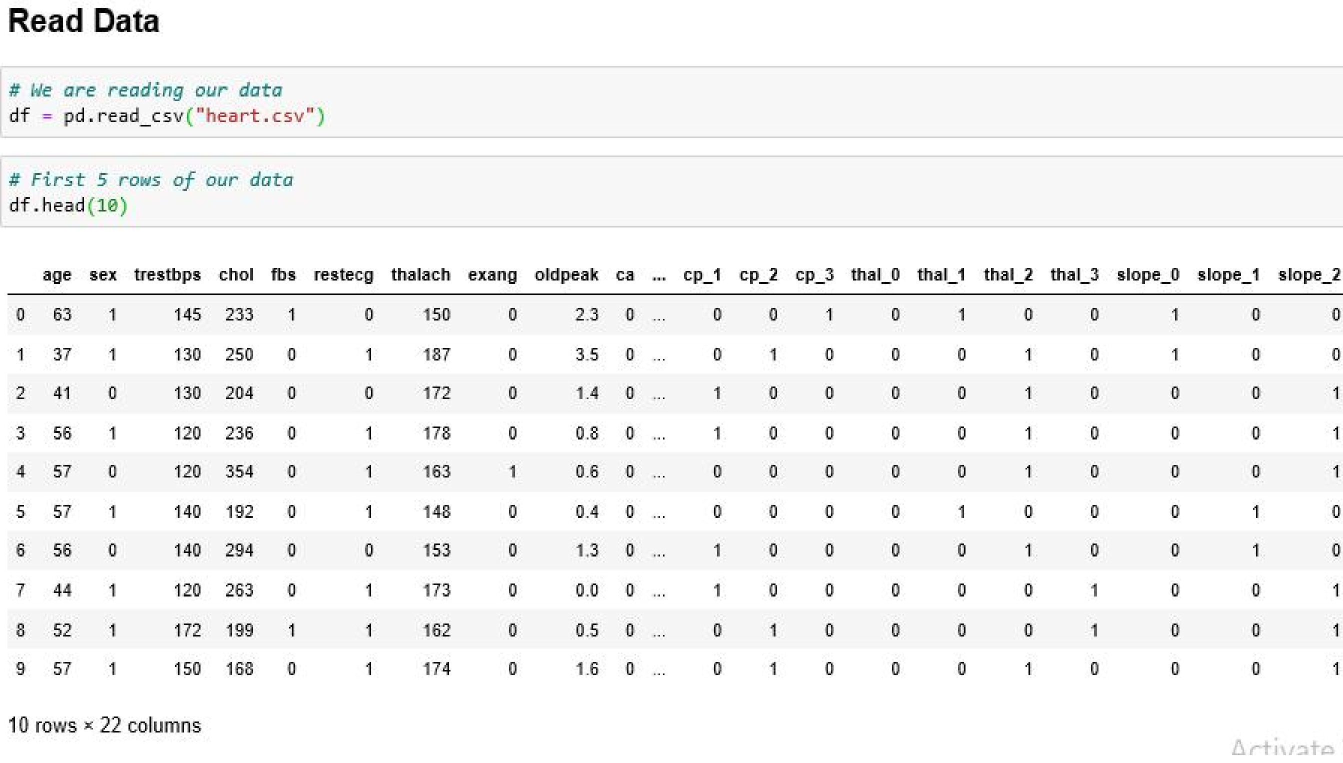
```
colors = ["purple", "green", "orange", "magenta","#CFC60E","#0FBBAE"]

sns.set_style("whitegrid")
plt.figure(figsize=(16,5))
plt.yticks(np.arange(0,100,10))
plt.ylabel("Accuracy %")
plt.xlabel("Algorithms")
sns.barplot(x=list(accuracies.keys()), y=list(accuracies.values()), palette=colors)
plt.show()
```

Confusion Matrixes



Data



Read Data

```
# We are reading our data
df = pd.read_csv("heart.csv")
```

```
# First 5 rows of our data
df.head(10)
```

	age	sex	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	ca	...	cp_1	cp_2	cp_3	thal_0	thal_1	thal_2	thal_3	slope_0	slope_1	slope_2
0	63	1	145	233	1	0	150	0	2.3	0	...	0	0	1	0	1	0	0	1	0	0
1	37	1	130	250	0	1	187	0	3.5	0	...	0	1	0	0	0	1	0	1	0	0
2	41	0	130	204	0	0	172	0	1.4	0	...	1	0	0	0	0	1	0	0	0	1
3	56	1	120	236	0	1	178	0	0.8	0	...	1	0	0	0	0	1	0	0	0	1
4	57	0	120	354	0	1	163	1	0.6	0	...	0	0	0	0	0	1	0	0	0	1
5	57	1	140	192	0	1	148	0	0.4	0	...	0	0	0	0	1	0	0	0	1	0
6	56	0	140	294	0	0	153	0	1.3	0	...	1	0	0	0	0	1	0	0	1	0
7	44	1	120	263	0	1	173	0	0.0	0	...	1	0	0	0	0	0	1	0	0	1
8	52	1	172	199	1	1	162	0	0.5	0	...	0	1	0	0	0	0	1	0	0	1
9	57	1	150	168	0	1	174	0	1.6	0	...	0	1	0	0	0	1	0	0	0	1

10 rows × 22 columns

References:

- A, A. S., & Naik, C. (2016). Different Data Mining Approaches for Predicting Heart Disease.
- Beyene, C., & Kamat, P. (2018). Survey on prediction and analysis the occurrence of heart disease using data mining techniques. International Journal of Pure and Applied Mathematics.
- Brownlee,J (2016) Naive Bayes for Machine Learning.
- Kirmani, M. (2017),Cardiovascular Disease Prediction using Data Mining Techniques.
- kaggle/python docker image: https://github.com/kaggle/docker-python
- Sigmoid function(https://www.geeksforgeeks.org/implement-sigmoid-function-using-numpy/)