Heart Disease prediction System using data mining and ML

Team Members Name

Abhinav Saurabh MT20127

Apoorv Lokhande MT20022

Shivam Mittal MT20041

INDRAPRASTHA INSTITUTE *of* INFORMATION TECHNOLOGY **DELHI**

Background Knowledge

- Heart disease is one of the most notable causes of death in the world today.
- Prediction of cardiovascular disease is regarded as one of the most important subjects in the section of clinical data analysis.
- One person dies every 36 seconds in the United States from cardiovascular disease*.
- Machine learning (ML) proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry.

*Reference: https://www.cdc.gov/heartdisease/facts.htm

Dataset

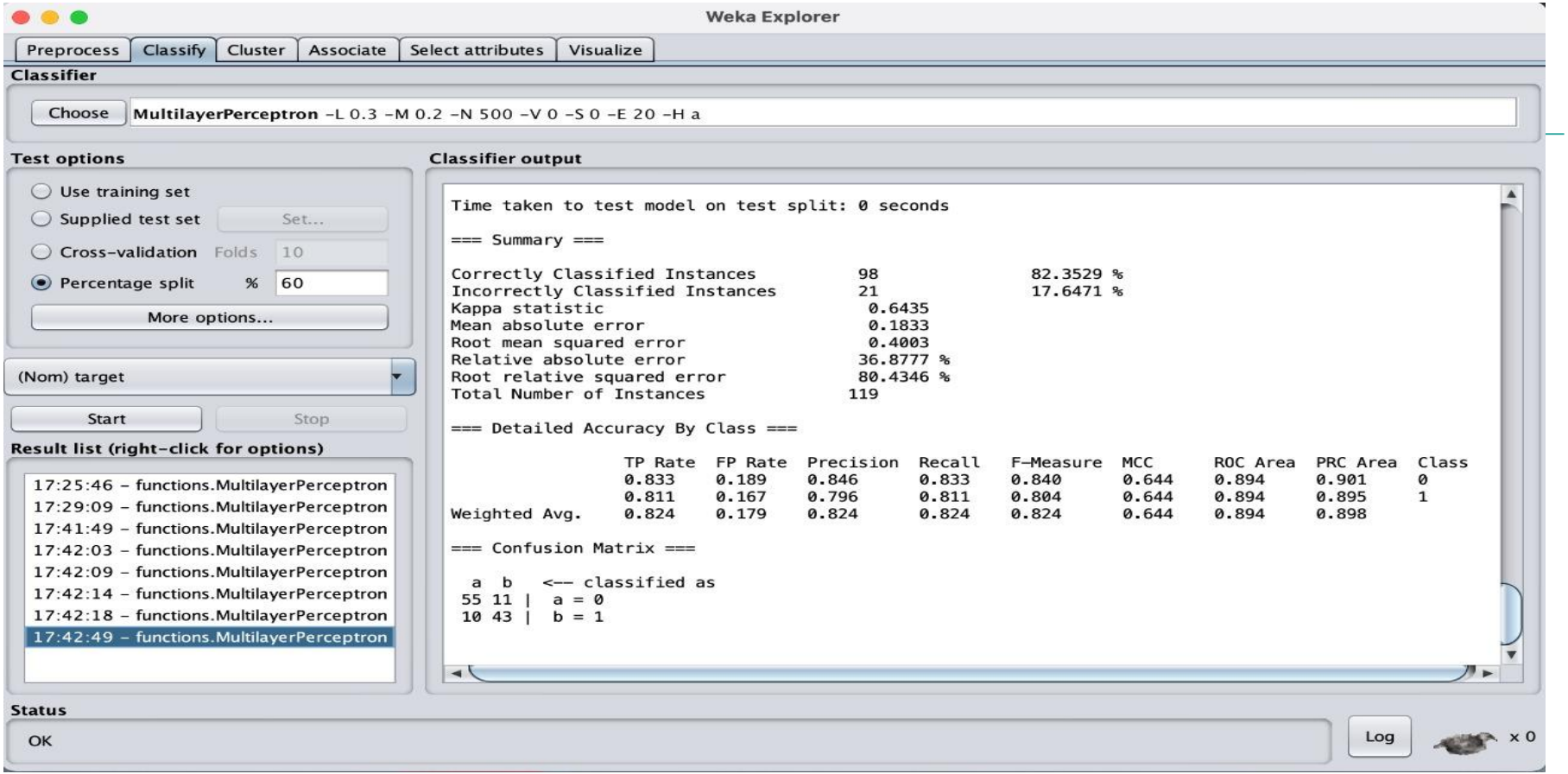
We have taken the heart disease dataset from the UCI Machine Learning repository. In the dataset there are 303 instance and 15 features present.

Features present in the dataset are age, sex, resting blood sugar, chest pain type, serum cholesterol, fasting blood sugar, resting electrographic, maximum heart rate, ST depression, exercise induced angina, ca, slope, defect type, target.

Dataset link - https://archive.ics.uci.edu/ml/datasets/heart+disease

Replication of the Work

- In the paper, authors have proposed an effective heart disease prediction system (EHDPS) developed using neural network for predicting the presence of heart disease.
- The system used 14 medical parameters such as age, sex, blood pressure, cholesterol etc. for prediction.
- They used Multilayer Perceptron NN with backpropagation algorithm for predicting the presence or absence of a heart disease.
- A Data mining tool Weka 3.6.11 was used for the experiments.



Accuracy obtained using Multilayer perceptron in weka is 82.35%

Present Work

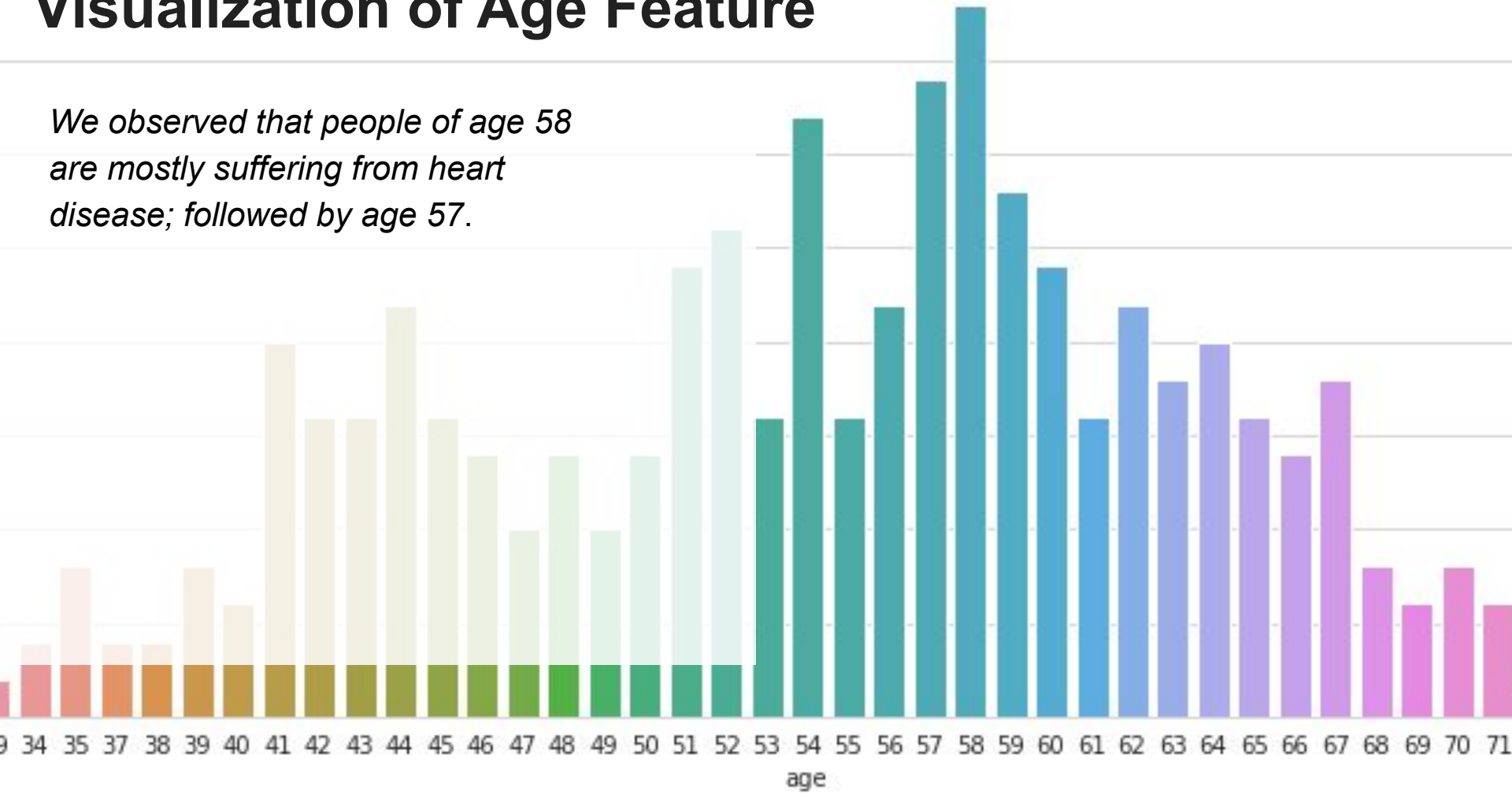
As mentioned, we have taken the cleveland heart disease dataset from UCI repository and then applied some basic preprocessing on the dataset as follows:

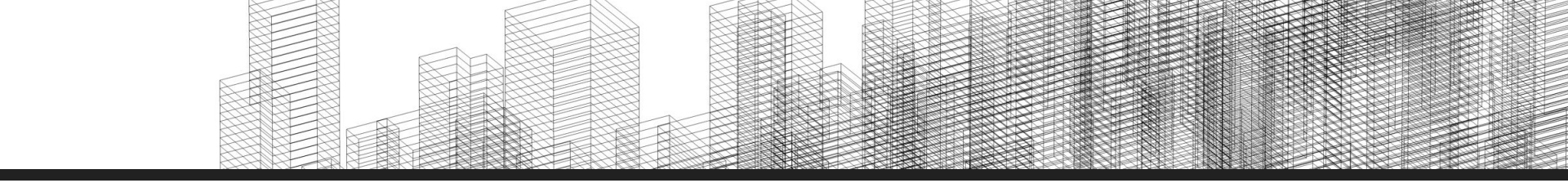
Cleaning and Preprocessing :

- Since the dataset contains some missing values in the form of ‘?’ in two columns *thalassemia* and *major_vessels_num*. So we have removed such rows
  from the dataset.
- Converted data types of some features into appropriate types. For instance, *age* is of type float, we converted it into int.

Visualization of Age Feature

We observed that people of age 58 are mostly suffering from heart disease; followed by age 57.

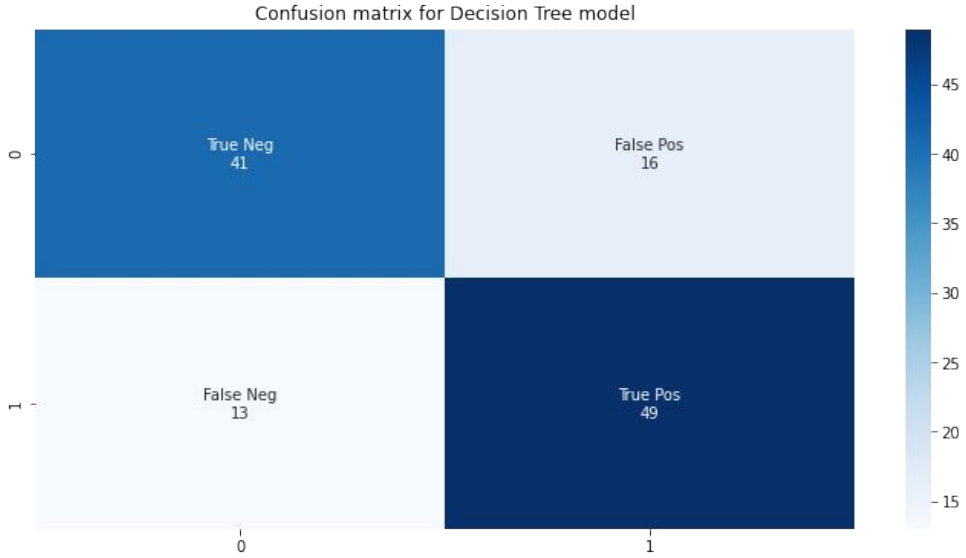




Applied Models

- Decision Tree
- Random Forest
- SVM
- K-Nearest Neighbour
- Multilayer Perceptron

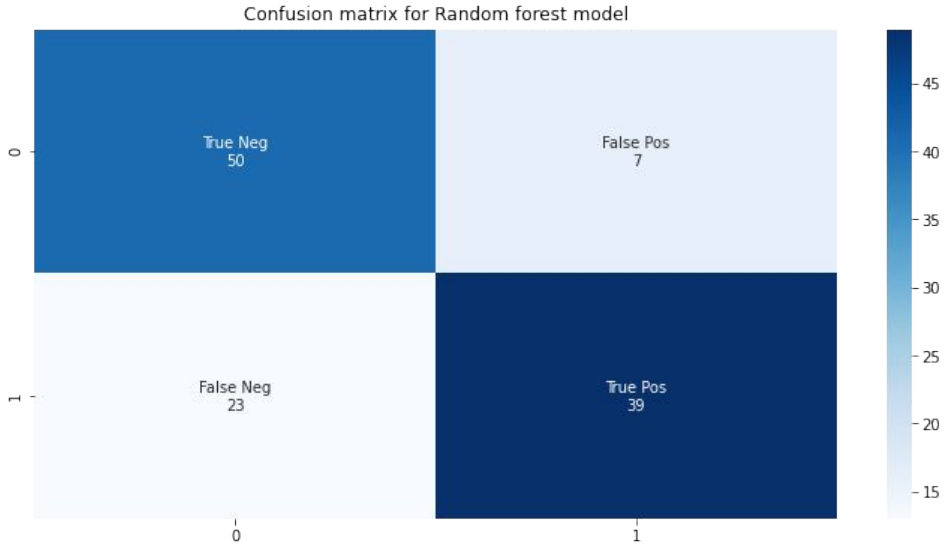
Decision Tree



- It is non-parametric supervised method.
- It is tree classifier structure
- It is used for classification as well as regression problems.
- Mostly preferred for solving classification problems.

- **In Decision Tree we obtained accuracy of 75.63%**
- False Positives = 16 , False Negative = 13
- We have used default parameters and will fine tune work later.

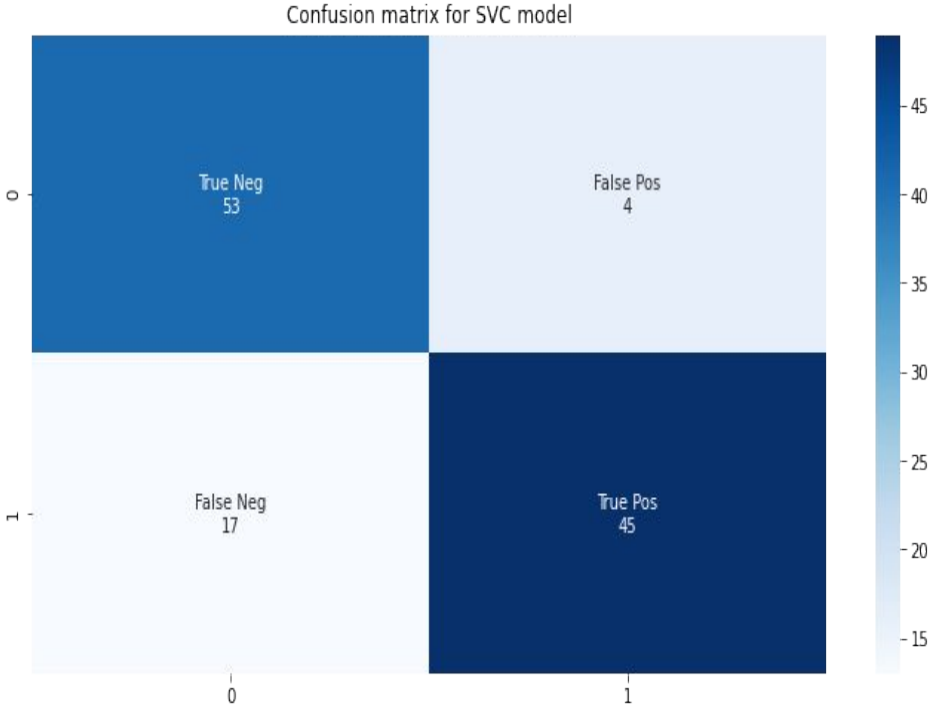
Random Forest



- Random forest is ensemble method.
- It is combination of various decision trees.
- It is used for classification as well as regression problems.
- Data characteristics can affect their performance.

- **In Random Forest we obtained accuracy of 74.38%**
- False Positives = 7, False Negative = 23
- We have used N estimator =10.
- We will optimize parameters in future work.

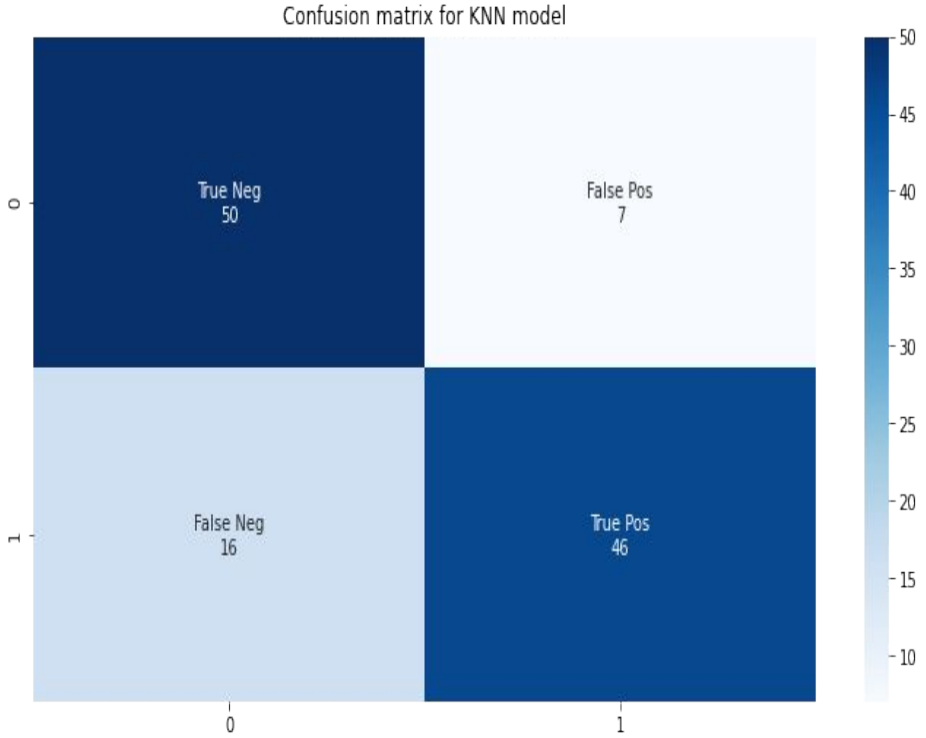
SVM



- SVM is supervised ML algorithm
- It is best used for classification problems.
- It separates the data points using hyperplane.

- **In SVM we obtained accuracy of 82.35%**
- False Positives = 4 , False Negative = 17
- We use rbf kernel.

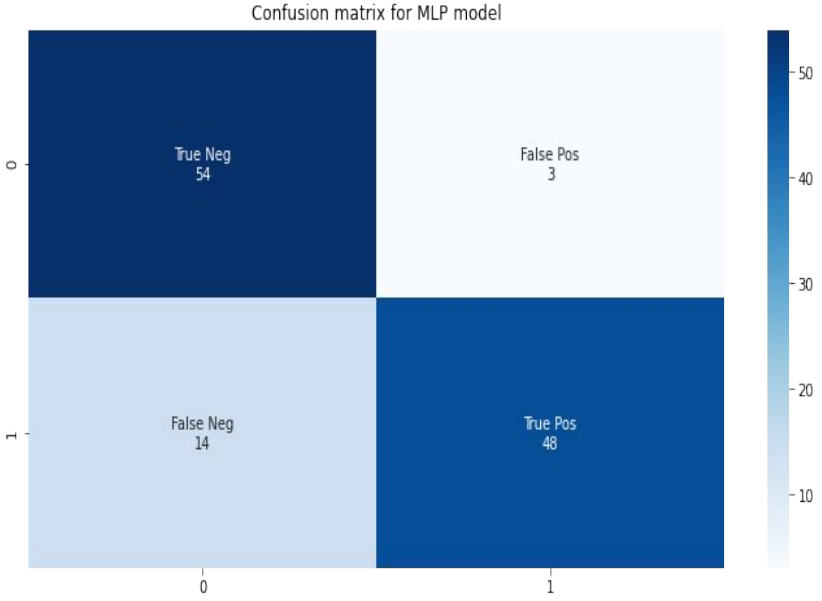
KNN



- KNN is supervise ML algorithm
- Used for both Classification and regression
- It uses feature similarity to predict values of new data points.

- **In KNN we obtained accuracy of 80.67%**
- False Positives = 7, False Negative = 16
- We have used neighbours=3

Multilayer perceptron



- It is an class of feedforward artificial neural networks.
- It contains at least 3 hidden neural networks i.e a input, a hidden and a output layer.
- It utilizes back propagation technique.

- In MLP we obtained accuracy of 85.71% .
- As seen from confusion matrix fewer False Positives and False Negatives
- False Positives = 3 , False Negative = 14
- **85.71% accuracy is highest among all.**
- We have used activation='tanh'
- For solver used sgd
- For learning rate used adaptive

Results

Models	Accuracy
Decision Tree	75.63 %
Random Forest	74.78 %
SVM	82.35 %
KNN	80.67 %
Multi layer Perceptron	**85.71 %**

Future Work

- Fine tuning our models; choosing suitable hyperparameters etc.

- Develop a UI in which a user can enter the details about him/her such as age, sex, cholestrol value and other features and then our model will generate predictions.