Heart Disease Prediction

By

Jay Ghodke

Amey Bargal

A project report submitted to the graduate faculty in partial fulfilment of the requirements of

IE 587 Big Data Analytics and Optimization class

Major: Industrial and Manufacturing Systems Engineering

Guided by

Dr. Guiping Hu, Faculty Professor

Iowa State University

Ames, Iowa

2019

Abstract

Heart diseases affect human life drastically. Heart diseases are the leading cause of human deaths worldwide. The ability of a heart to perform its dedicated pumping of blood to all the parts of the body is hampered in a heart disease. Many times the symptoms are not easily visible which make heart diseases more risky. Accurate and early diagnosis of a heart disease is very crucial in preventing further effects of the heart disease. Many factors contribute to the cause of heart disease. Therefore, to classify people with a risk of heart disease and healthy people, machine-learning methods can be used which prove to be dependable and efficient. In this study, we have used heart disease dataset of the people of Framingham to build a machine-learning based prediction model. We used four machine-learning algorithms, one feature selection package, cross validation method and evaluated the performance of the classifiers based on classification accuracy. The proposed model can be used to help the doctors to diagnose a person noninvasively for heart diseases.

Chapter 1 – Introduction

Heart diseases are the number one cause of death globally, taking an estimated 17.9 million lives every year. The increase intake of fast food and junk food in the daily appetite, excessive stress, hereditary medical history, hectic routines and lack of daily exercise have majorly contributed towards rapid growth of cardiovascular diseases or heart diseases. The symptoms of heart disease include shortness of breath, weakness of physical body, swollen feet, and fatigue with related signs, for example, elevated jugular venous pressure and peripheral edema caused by functional cardiac or non-cardiac abnormalities. Diagnosis of heart diseases in early stages is very intricate. Moreover, lack of experienced cardiologists and apparatus in many countries affect the rate of heart diseases. Thus, accurate and on-time diagnosis of heart diseases is essential for reducing the risk of severe heart problems and complexity in solving them.

The present diagnosis techniques include Electrocardiogram (ECG), Stress test, 2D Echo test, heart MRI, angiography, etc. These techniques can give false results due to human errors and the cost and time factor is another reason to worry.

Noninvasive based methods which include machine learning algorithms and models prove to be reliable and more efficient than the previously mentioned techniques. Predictive algorithms like Logistic Regression, Decision Trees, Support Vector Machines and k-Nearest Neighbors were used in our machine learning based model to classify the possible heart patients and healthy people. These algorithms are used by many researchers and have proven to be accurate with very less execution time as compared to the traditional diagnostics. The proposed machine learning model can be used by the cardiologists and physicians to run preliminary diagnostic tests on the patient to get an overview on their heart health status.

Chapter 2 – Problem Statement

2.1 Analyze the most relevant factors of heart disease

The dataset used by us to train the model has 15 attributes with over four thousand observations. In order to obtain maximum accuracy, it is necessary to identify the factors which contribute the most towards determining the probability of causing a heart disease.

2.2 Predict the overall risk of heart disease

After determining the most relevant factors, the risk of a heart disease in future for given details of a person is calculated in the term of binary output, that is '1' being possible development of cardiovascular disease in ten years and '0' being healthy.

2.3 Dataset

The dataset is publically available on the Kaggle website, and it is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The dataset includes more than four thousand observations and 15 features. Every feature is chosen as potential risk factor. There are both demographic medical risk factors in the dataset.

Features -

<u>Demographic:</u>

Sex: male or female (Nominal)

Age: Age of the patient (Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)

Education: no further information provided

Behavioral:

Current Smoker: whether or not the patient is a current smoker (Nominal)

Cigarettes Per Day: the number of cigarettes that the person smoked on average in one day.(can be considered continuous as one can have any number of cigarettes, even half a cigarette.)

Information on medical history:

BP Meds: whether or not the patient was on blood pressure medication (Nominal)

Prevalent Stroke: whether or not the patient had previously had a stroke (Nominal)

Prevalent Hypertension: whether or not the patient was hypertensive (Nominal)

Diabetes: whether or not the patient had diabetes (Nominal)

Information on current medical condition:

Tot Cholesterol: total cholesterol level (Continuous)

Systolic BP: systolic blood pressure (Continuous)

Diastolic BP: diastolic blood pressure (Continuous)

BMI: Body Mass Index (Continuous)

Heart Rate: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)

Glucose: glucose level (Continuous)

Target variable to predict:

10 year risk of coronary heart disease (CHD) - (binary: “1”, means “Yes”, “0” means “No”)

male	age	education	rrentSmok	igsPerDay	BPMeds	evalentStro	revalentHy	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
1	39	4	0	0	0	0	0	0	195	106	70	27	80	77	0
0	46	2	0	0	0	0	0	0	250	121	81	28.7	95	76	0
1	48	1	1	20	0	0	0	0	245	127.5	80	25.3	75	70	0
0	61	3	1	30	0	0	1	0	225	150	95	28.6	65	103	1
0	46	3	1	23	0	0	0	0	285	130	84	23.1	85	85	0
0	43	2	0	0	0	0	1	0	228	180	110	30.3	77	99	0
0	63	1	0	0	0	0	0	0	205	138	71	33.1	60	85	1
0	45	2	1	20	0	0	0	0	313	100	71	21.7	79	78	0
1	52	1	0	0	0	0	1	0	260	141.5	89	26.4	76	79	0
1	43	1	1	30	0	0	1	0	225	162	107	23.6	93	88	0
0	50	1	0	0	0	0	0	0	254	133	76	22.9	75	76	0
0	43	2	0	0	0	0	0	0	247	131	88	27.6	72	61	0
1	46	1	1	15	0	0	1	0	294	142	94	26.3	98	64	0

Table 1: Dataset

Chapter 3 – Methodology for the proposed model

The proposed model has been developed with the goal to classify people with heart disease and healthy people. The performances of different machine learning predictive models for heart disease diagnosis on selected features were tested. Feature selection algorithm such as SelectKBest was used to select most important features out of 15 features.

The methodology of the proposed system structured into five stages including –

1. Preprocessing of dataset
2. Data exploration
3. Feature selection
4. Feature scaling
5. Data splitting
6. Modelling
7. Evaluation and results

3.1 Preprocessing of dataset

In this step, we initially investigated the dataset for any missing or duplicate values. Preprocessing of data is crucial for effective implementation of the machine learning model to give better output. All the missing values were deleted from the dataset along with the duplicate values. All these steps were used for the preprocessing of the data.

3.2 Data exploration

In data exploration, the distribution of all features is explored by plotting histograms for each of them.

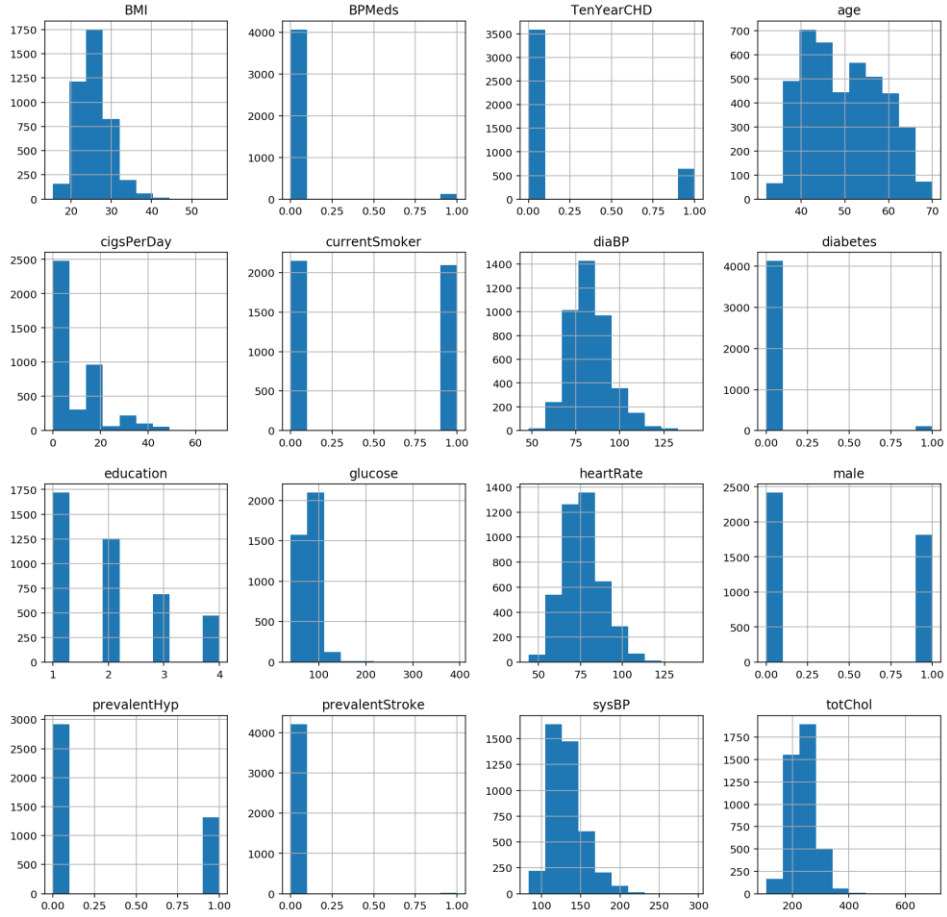

Figure 1: Histogram

Moreover, to check the correlation with other features and the outcome variable, a correlation plot is used as a better graphical representation of their correlation. From the correlation it is evident that the feature ‘education’ does not contribute towards the cause of heart diseases as level education cannot determine if a person will or will not have heart related issues. Also, systolic BP, Age, total Cholesterol and glucose are found to most correlated with the outcome variable.

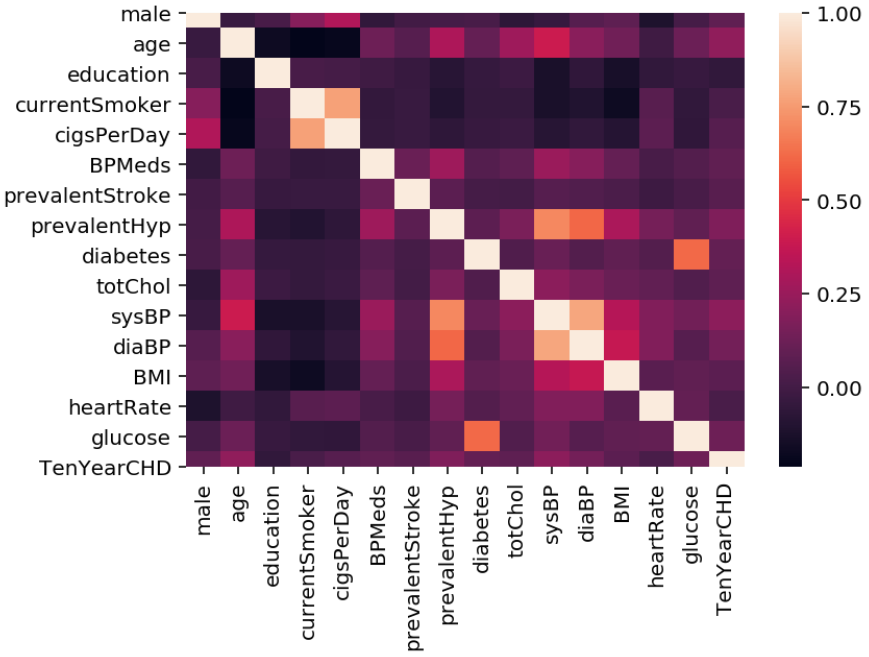

Figure 2: Correlation Plot

3.3 Feature selection

In our dataset, various irrelevant features are present which have negligible contribution towards the outcome variable. From the correlation plot we found out the ‘education’ was highly irrelevant. Feature selection is a very important step in building a machine learning model because it improves the overall accuracy and the model requires less time for execution as well. We used SelectKBest algorithm from the Scikit Learn package to select 10 best features. Therefore, we will only keep those features that have the strongest relationship with the output variable. These features are

Systolic Blood Pressure, Glucose, Age, Cholesterol, Cigarettes per Day, Diastolic Blood Pressure, Hypertension, Diabetes, Blood Pressure Medication and Gender

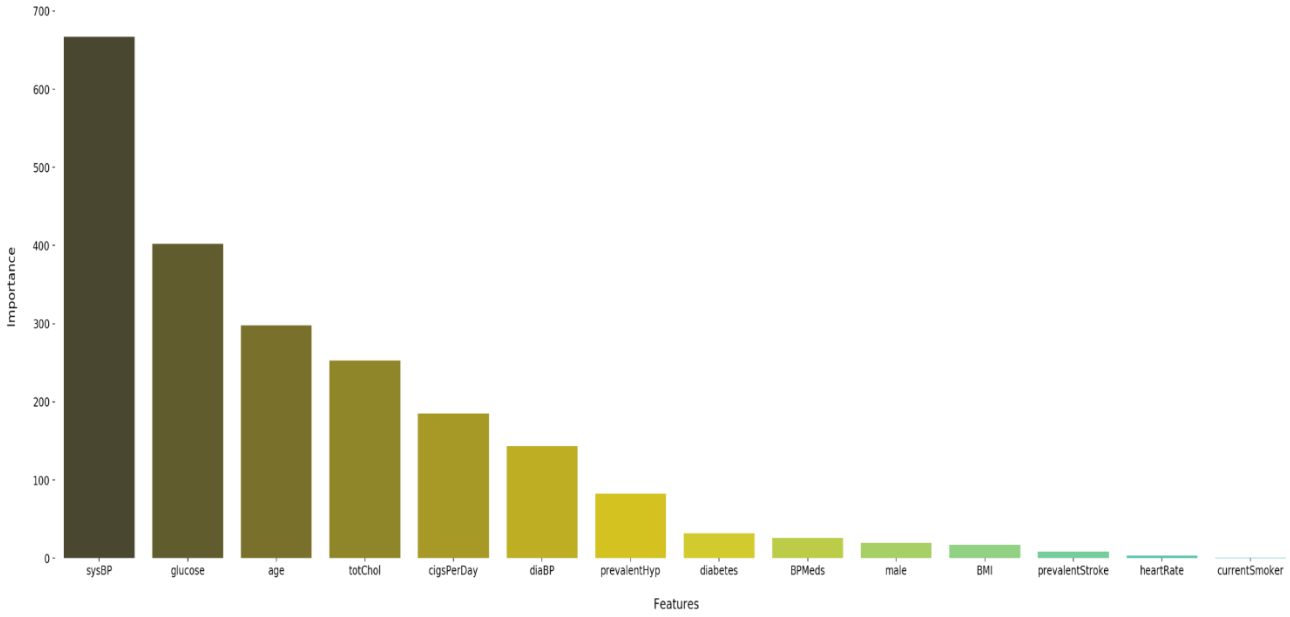

Figure 3: Correlation Plot

3.3.1 Finding outlier in the features

Outliers are the observations which significantly deviate from the other group of observations. If not removed from the data, they can negatively impact the accuracy of the model.

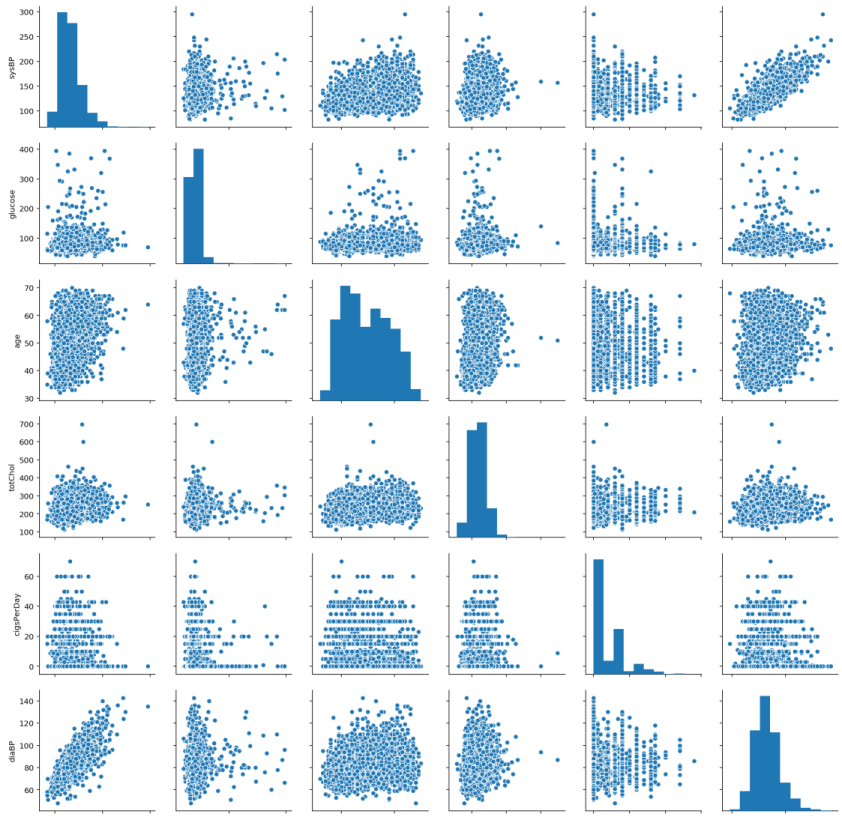

Figure 4: Correlation Plot

From the pair-plot it can be seen that ‘cholesterol’ feature has two observation deviating drastically from the rest of observations. Hence, these two outliers are removed from the dataset.

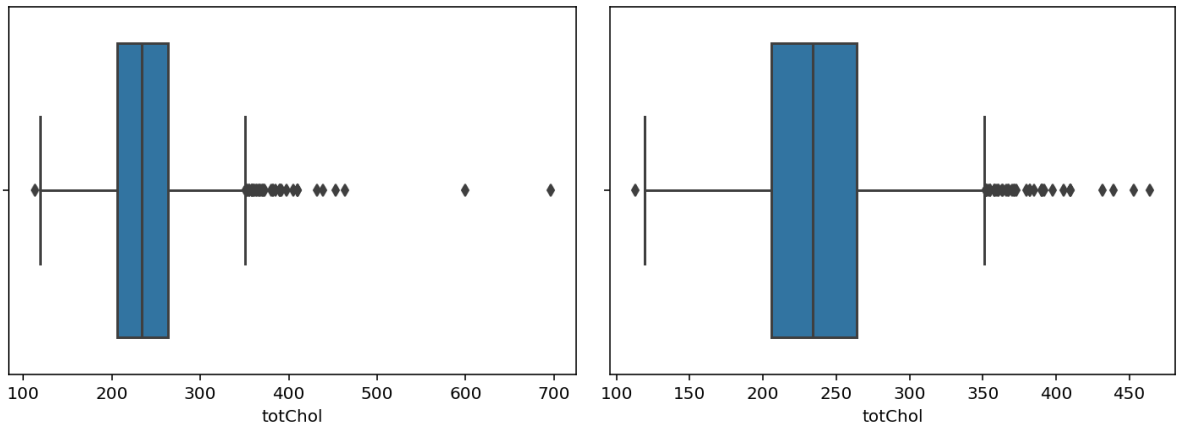

Figure 5: Box Plot for Cholesterol before and after removing outliers

3.4 Feature Scaling

Feature scaling is essential in any machine learning model since we want to try out different models, and also these that use distance as a measure. Thus, normalizing the dataset will help us use various classifiers on the dataset without any changes in the data.

3.5 Data splitting

The dataset was partitioned into training and test sets in the ratio of 80% and 20% respectively.

3.5.1 Resampling imbalanced dataset

The count of people not having heart disease compared to the patients having heart diseases was highly imbalanced with a ratio of 3178 : 571 or 5.57 : 1 respectively. Using a highly imbalanced dataset can be false leading, the classifier will always predict the “most common” class without performing much analysis on other features. The model will show higher accuracy but it will misleading.

- Under-sampling the imbalanced dataset

In this process we decreased the number of observations from the majority class to level it with the other feature. If we do not balance the number of observations, most classification algorithms will mostly focus on the majority class. Random observations from the majority class were removed to achieve a balanced dataset.

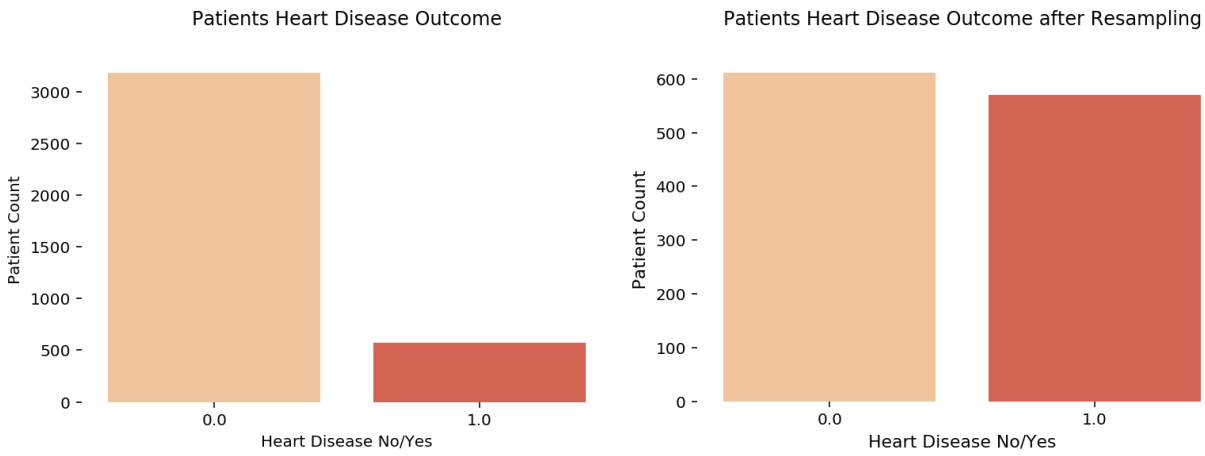

Figure 6: Patients count

3.6 Modelling

1. Logistic Regression

   Logistic regression is a machine learning based classifier. It is mostly used when the response in binary.

   Accuracy obtained by LR is 68.3%

   F1 score is 43.1% *(The F1 score can be interpreted as a weighted average of the precision and recall)*

   Precision score 30.4% *(When it predicts yes, how often is it correct? Precision=True Positive/predicted yes)*

   Recall score 73.8% *(When it's actually yes, how often does it predict yes? True Positive Rate = True Positive/actual yes)*

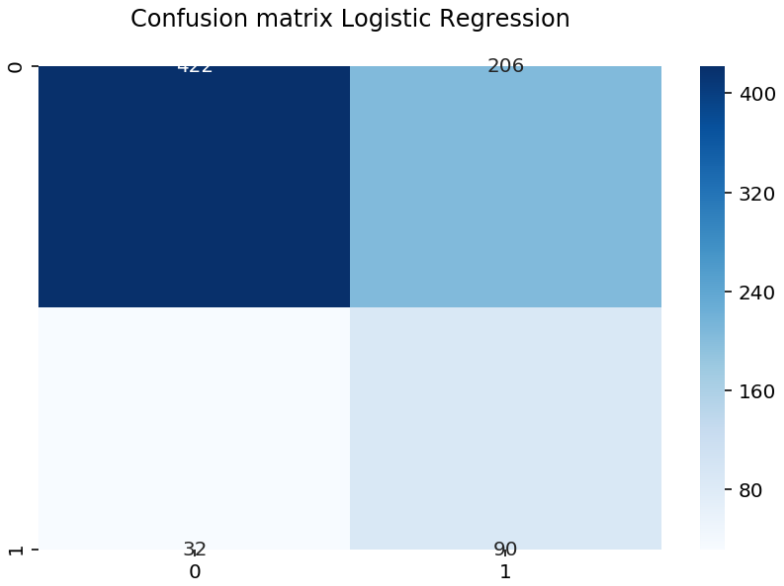

Figure 7: Confusion matrix

2. Support Vector Machine (SVM)

The SVM is a machine learning classification algorithm which has been mostly used for classification problems. Four types of kernels were tested for the SVM model.

Accuracy obtained by SVM is 78.5% for polynomial kernel. The accuracy for linear, radial and sigmoid kernel is 66%, 67% and 69% respectively.

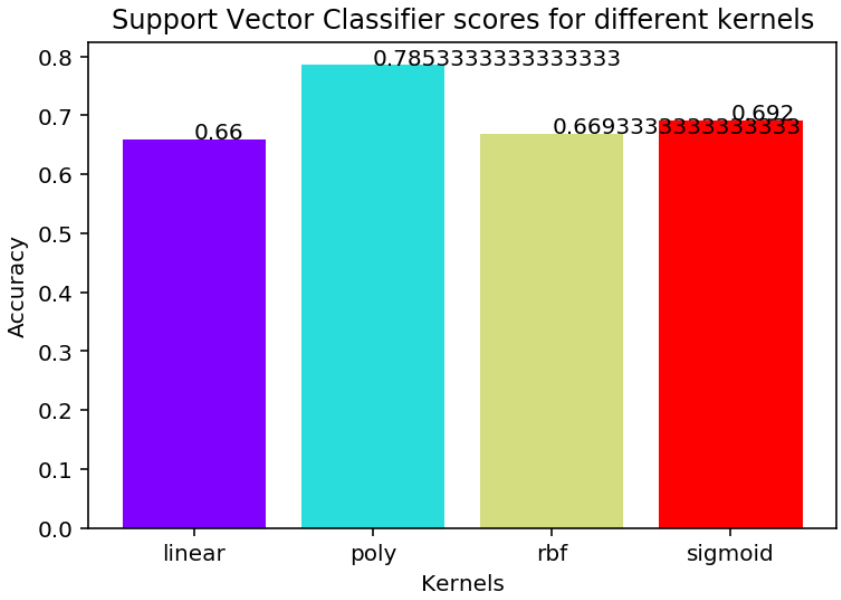

Figure 8: SVM kernel vs. Accuracy plot

3. Decision Trees

A decision tree is a supervised machine learning algorithm. The techniques of the decision tree are simple and easily understandable for how to take the decision.

Accuracy obtained by DT is 74.2%

F1 score is 53% *(The F1 score can be interpreted as a weighted average of the precision and recall)*

Precision score 36.1% *(When it predicts yes, how often is it correct? Precision=True Positive/predicted yes)*

Recall score 100% *(When it's actually yes, how often does it predict yes? True Positive Rate = True Positive/actual yes)*

The decision tree classifier was tested on maximum number of features varying from 1 to 10. The maximum accuracy was for 6 maximum features.

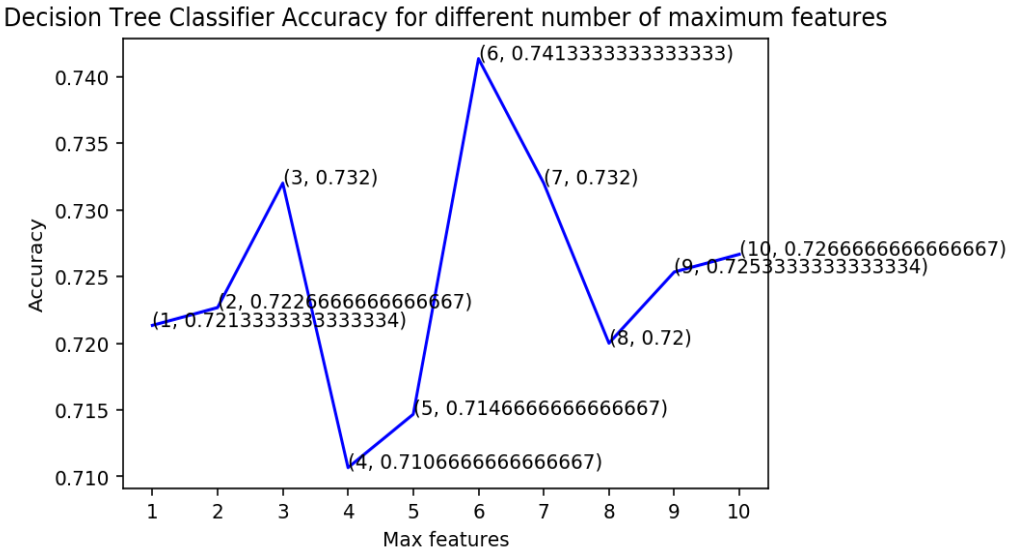

Figure 9: Decision Trees Classifier accuracy

4. k-Nearest Neighbors

K-NN is a supervised learning classification algorithm. K-NN algorithm predicts the class label of a new input; K-NN utilizes the similarity of new input to its inputs samples in the training set.

Accuracy obtained by KNN is 81.7%

F1 score is 50.2% *(The F1 score can be interpreted as a weighted average of the precision and recall)*

Precision score 45.1% *(When it predicts yes, how often is it correct? Precision=True Positive/predicted yes)*

Recall score 56.6% *(When it's actually yes, how often does it predict yes? True Positive Rate = True Positive/actual yes)*

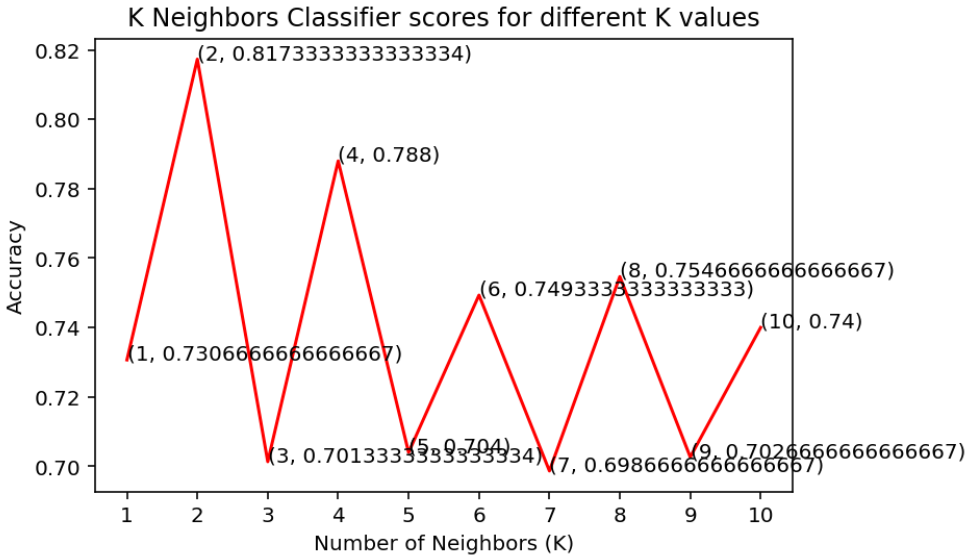

Figure 10: KNN Classifier accuracy

Accuracy after cross-validation is 83.7% where k = 6 in K – fold cross validation.

ROC Curve - The AUC ROC Curve is a measure of performance based on plotting the true positive and false positive rate and calculating the area under that curve. The closer the score to 1 the better the algorithm's ability to distinguish between the two outcome classes.

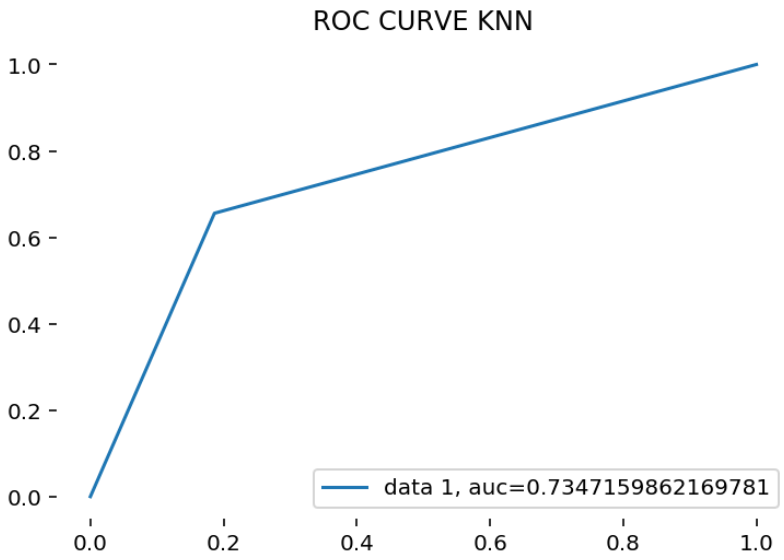

Figure 11: ROC curve for KNN

Chapter 4 - Results and Evaluation

Applying the model using KNN prediction classifier as the KNN model gave the highest classification accuracy with cross validation. We used confusion matrix to check the true positive rates for every classifier. The accuracy score for training and test sets were found to be similar, which means the KNN model did not over-fit.

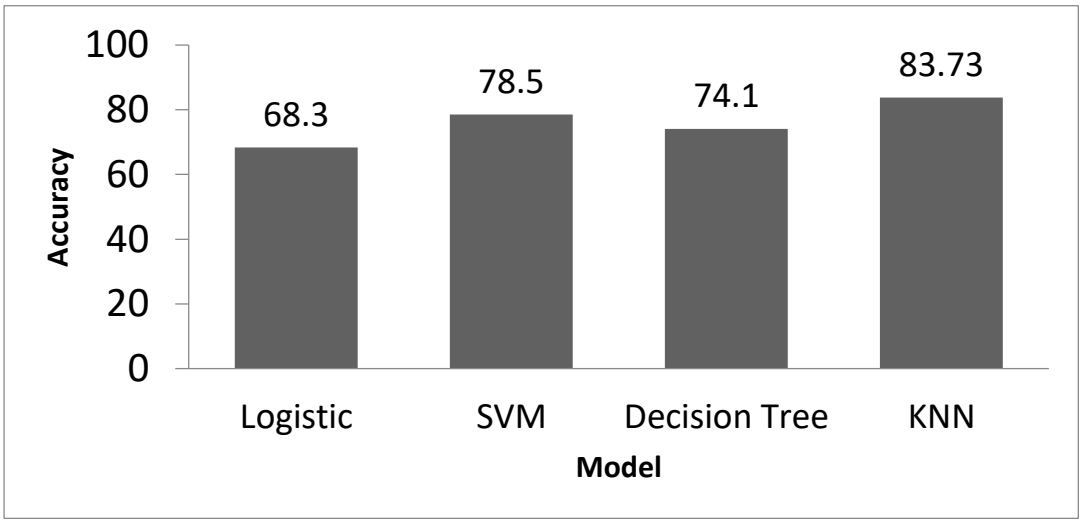

Figure 12: Accuracy for all model

```
Input Patient Information:
Patient's age: >>> 23
Patient's gender. male=1, female=0: >>> 1
Patient's smoked cigarettes per day: >>> 3
Patient's systolic blood pressure: >>> 125
Patient's diastolic blood pressure: >>> 85
Patient's cholesterin level: >>> 110
Was Patient hypertensive? Yes=1, No=0 >>> 0
Did Patient have diabetes? Yes=1, No=0 >>> 0
What is the Patient's glucose level? >>> 75
Has Patient been on Blood Pressure Medication? Yes=1, No=0 >>> 0


Result:
The patient will not develop a Heart Disease.
```

Figure 13: Questionnaire

Chapter 5 – Conclusions

In this project, a machine-learning-based predictive system was proposed for the diagnosis of heart disease. The system was tested on Framingham heart disease dataset. Four well-known classifiers such as Logistic Regression, K-NN, SVM and DT were used. Scikit learn's SelectKBest was used to decide top ten relevant features. The dataset was balanced to get better accuracy. K fold cross validation was used in KNN model to achieve more accuracy which is 83.7%. Due to the highest performance of KNN classifier, it was selected to predict the outcome of test set.

References

1. S. Palaniappan and R. Awang, “Intelligent heart disease prediction system using data mining techniques,” in Proceedings of IEEE/ACS International Conference on Computer Systems and Applications (AICCSA 2008), pp. 108–115, Doha, Qatar, March-April 2008
2. P. A. Heidenreich, J. G. Trogdon, O. A. Khavjou et al., “Forecasting the future of cardiovascular disease in the United States: a policy statement from the American Heart Association,” Circulation, vol. 123, no. 8, pp. 933–944, 2011
3. A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms, Mobile Information Systems, Volume 2018, Article ID 3860146, 21 pages
4. An Introduction to Statistical Learning, 2013