Cardiovascular Disease Prediction with Machine Learning

Khuldeena W. Najeeb[1] & Dr. M. S. Panse[2]

[1]Student M.Tech Electronics ,

[2]Professor, Department of Electronics Engineering, VJTI Mumbai

Abstract

Cardiovascular diseases are more than common these days and are on an increase day by day mainly due to lifestyle changes hence life of people is at risk. Each person has different values for Blood Pressure, Cholesterol and pulse rate. Medically proven normal values of Blood pressure is 120/90, Cholesterol is 100-129 mg/dL ,Pulse rate is 72, Fasting Blood Sugar level is 100 mg/dL ,Heart rate is 60-100 bpm, ECG is normal, Width of major vessels is 25 mm (1 inch) in the aorta to only 8 μm in the capillaries. This paper surveys different classification techniques used for predicting the risk level of each person based on gender, Blood Pressure, gender, cholesterol, pulse rate. This Prediction system based on predictive modelling predicts the disease of the user on the basis of the symptoms that user provides as an input to the system. The prediction system analyses the symptoms provided by the user as input and gives the probability of the disease as an output. Disease Prediction is done through implementing 4 different techniques namely Naïve Bayes, KNN, Decision Tree and Linear Regression Algorithms. These techniques calculate the probability of the disease. Hence, average prediction accuracy is obtained with probability 83%.

Keywords: CVD, Blood Pressure, Cholesterol, Naives Bayesian, KNN, Decision Tree, Linear Regression

1. Introduction

A lot of factors affect the cardiovascular system of a human, a lot of problems are occurring at a rapid pace. Human heart, being an essential organ in a human body which pumps blood through the body for the blood circulation is essential and its health is to be conserved for a healthy living. The increase in the number heart disease patients may be due to the bad eating habits, lack of sleep, restless nature, depression and numerous other factors such as obesity, poor diet, family history, high blood pressure, high blood cholesterol, idle behavior, family history, smoking and hypertension.

Healthcare industry has been generating huge amounts of data about patients and their disease diagnosis reports are being especially taken for the prediction of heart attacks worldwide. Since the data about heart disease is colossal, the machine learning techniques can be implemented for the analysis. In this paper Data mining will be used viz. a task of extracting the vital decision-making information from a collective of past records for future analysis or prediction. Therefore by using medical data mining it is possible to provide insights on a patient's history and is able to provide clinical support through the analysis.

Implementation, training and testing of the modules have been done on datasets obtained from UCI and Physio net data repositories. Data format have been modified in case of the ECG report data for betterment of action by the convolutional neural network used in the

research and in the risk prediction module we have chosen attributes for training and implementing the multi-layered neural network developed.

2. Methodology

2.1. Dataset –
In this paper 4 algorithms have been used to find out the reasons of heart disease and create a model to get the maximum accuracy possible. For this we have used UCI dataset[6] where 4 databases: Cleveland, Hungary, Switzerland, and the VA Long Beach are already stored and the Creators are:
Heart Disease Prediction using Machine Learning Algorithms 49
1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D. And the Donor is: David W. Aha (aha '@' ics.uci.edu) (714) 856-8779. From this dataset 5 widely used algorithms Logical Regression, Random forest, K-Nearest Neighbors and Decision Tree will be implemented to create the model with the maximum accuracy possible.

2.2. Software-

One Software is used in this Project- Python Spyder

2.2.1. Python Spyder

Spyder is a powerful scientific environment written in Python, for Python, and designed by and for scientists, engineers and data analysts. It offers a unique combination of the advanced editing, analysis, debugging, and profiling functionality of a comprehensive development tool with the data exploration, interactive execution, deep inspection, and beautiful visualization capabilities of a scientific package. Beyond its many built-in features, its abilities can be extended even further via its plugin system and API.

3. Project Definition

The main goal of this paper is to predict cardiovascular disease occurrence with the highest accuracy. In order to achieve this, we will test some classification algorithms namely Logistic Regression, Naives Bayesian, KNN and Decision Tree. It includes all results obtained from the study and introduces the best performer according to accuracy metric. Several algorithms have been chosen typically for solving supervised learning problems throughout classification methods.

6. Results

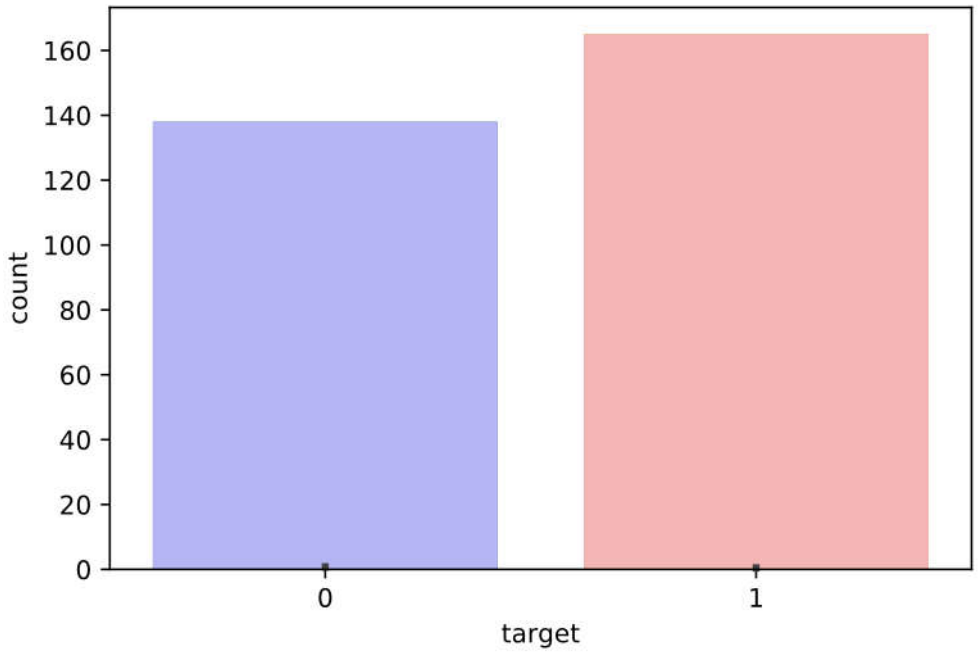


Figure 1. Percentage of Patients With or Without Heart Problems

This output depicts 0 as patients without heart disease i.e. 45.54% and 1 as patients with heart disease i.e. 55.46%

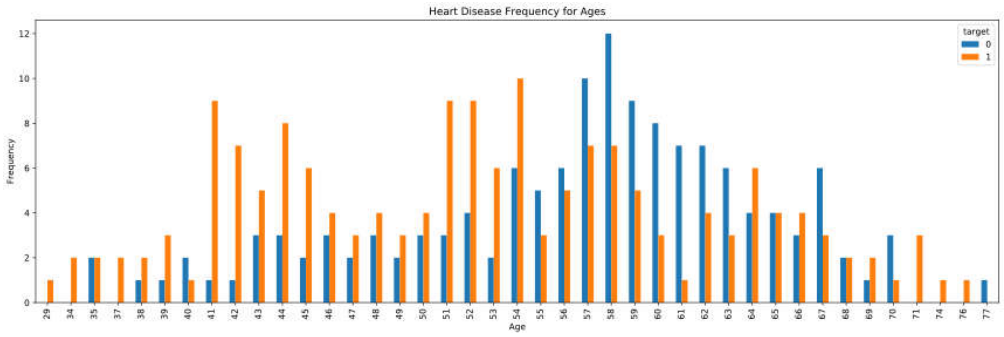


Figure 2. Heart Disease Frequency According to Age

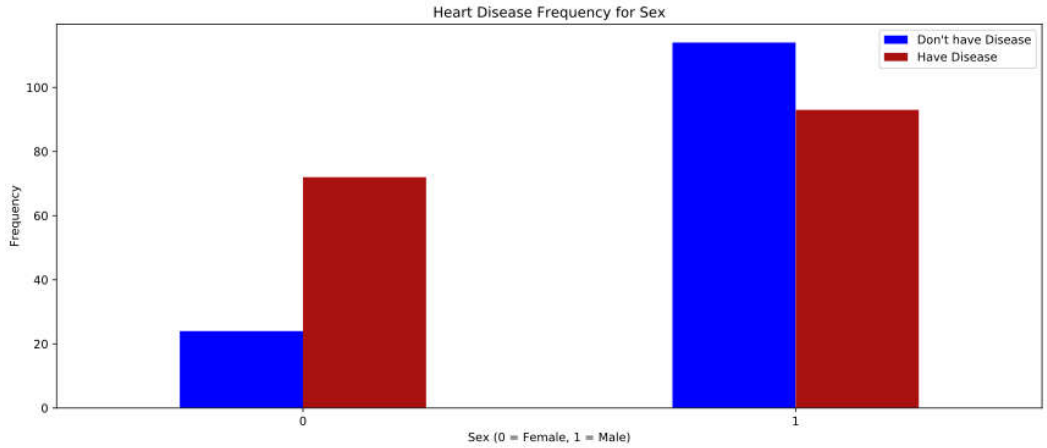


Figure 3. Heart Disease Frequency for Male and Female

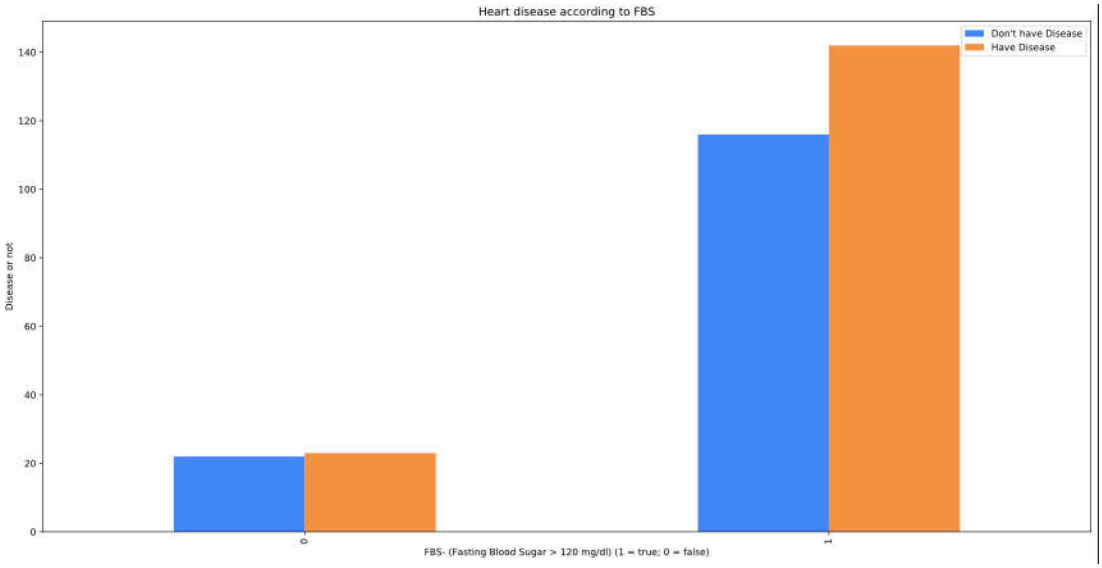


Figure 4. Fasting Blood Sugar Data

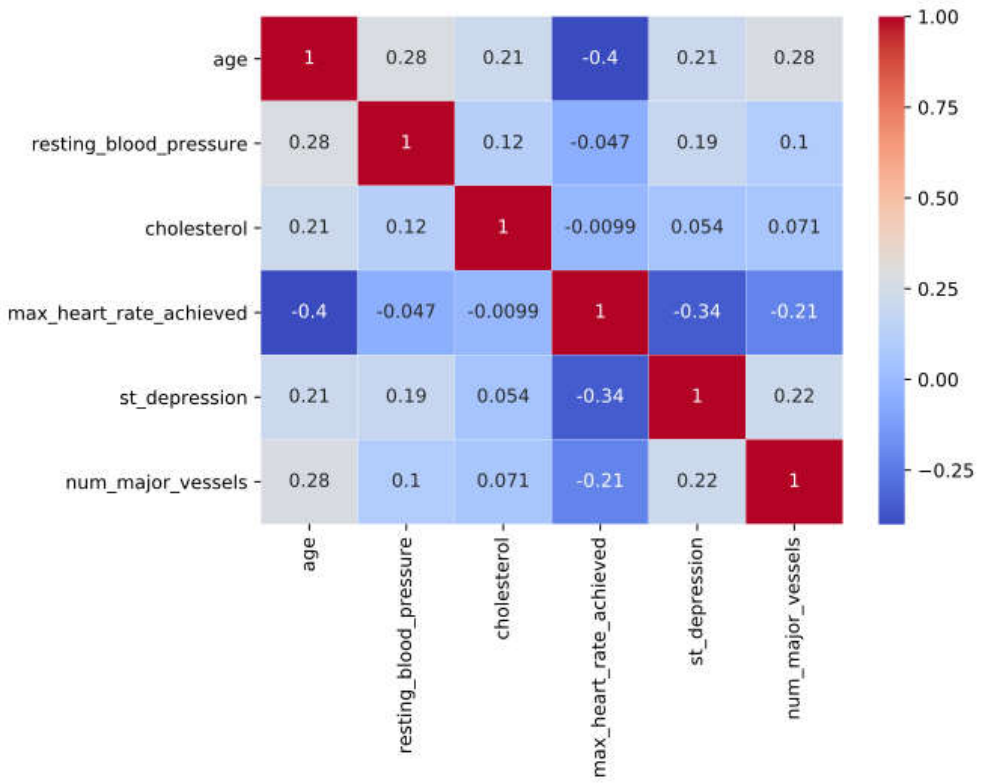
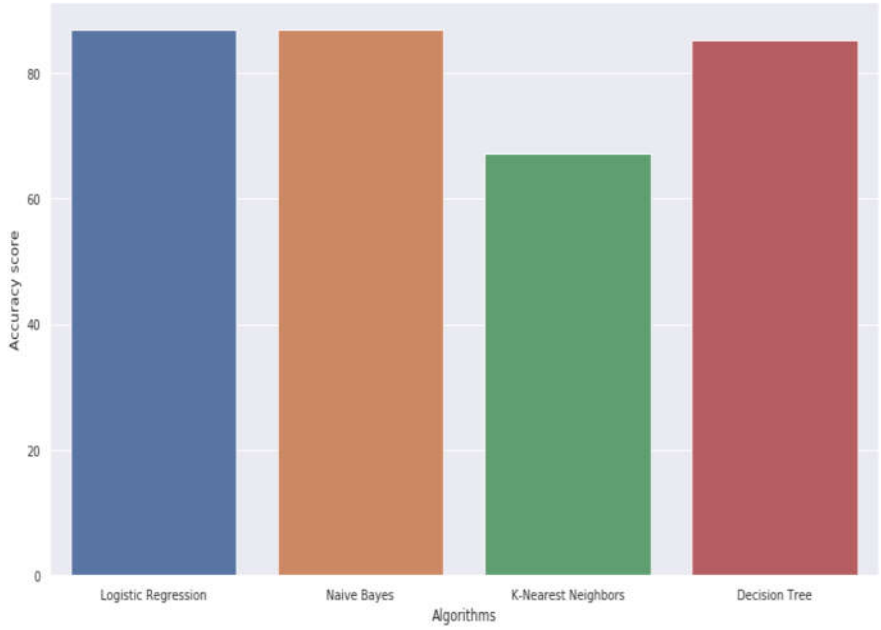


Figure 5. Correlation Plot




```
The accuracy score achieved using Logistic Regression is: 86.89 %
The accuracy score achieved using Naive Bayes is: 86.88524590163935 %
The accuracy score achieved using K-Nearest Neighbors is: 67.21 %
The accuracy score achieved using Decision Tree is: 85.24590163934427 %
```

Figure 6.1,6.2. Final Result

It is observed that, we can achieve the best accuracy of 86.89% using Logistic Regression.

7. Conclusion

This paper is presented using data mining techniques. Logistic regression, KNN, Naive Bayes, Decision Tree are used to develop the disease prediction system. Logistic Regression proves with better results and assists the domain experts and even the person related to the medical field to plan for a better and early diagnosis for the patient.

8. References

[1] M. K. Awang and F. Siraj, "Utilization of an artificial neural network in the prediction of heart disease," Int. J. Bio-Science Bio-Technology, vol. 5, no. 4, pp. 159–165, 2013.

[2] P. Selvakumar and S. P. Rajagopalan, "A survey on neural network models for heart disease prediction," J. Theor. Appl. Inf. Technol., vol. 67, no. 2, pp. 485–497, 2014.

[3] A. Methaila, P. Kansal, H. Arya, and P. Kumar, "Early Heart Disease Prediction Using Data Mining Techniques," vol. 6956, no. October, pp. 53–59, 2014.

[4] I. S. F. Dessai, "Intelligent Heart Disease Prediction System Using Probabilistic Neural Network," no. 5, pp. 38–44, 2013.

[5] T. Karayilan and Ö. Kiliç, "Prediction of Heart disease using neural network," in 2nd International Conference on Computer Science and Engineering, UBMK 2017, 2017, pp. 719–723.

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