Heart Disease Prediction Using Machine Learning

Prof. Vikarant A. Agaskar, Karan Kalla, Shubham Malankar

[1]*Assi.Prof. Vikrant A. Agaskar, Dept of Computer Engineering, Vidhyavardhini's College of Engineering and Technology, Vasai(W), Maharashtra, India.*
[2]*Karan Kalla, Dept of Computer Engineering, Vidhyavardhini's College of Engineering and Technology, Vasai(W), Maharashtra, India.*
[3]*Shubham Malankar, Dept of Computer Engineering, Vidhyavardhini's College of Engineering and Technology, Vasai(W), Maharashtra, India.*

Abstract - *An Heart diseases cause a high mortality rate in the world. Prediction and diagnosing of heart disease has become a challenging task faced by doctors and hospitals both in India and abroad. Heart Disease Prediction System is the system that helps to predict the heart disease mainly cardiovascular disease that includes Myocardial infractions. Data mining techniques and machine learning algorithms play a very important role in this area. The researchers accelerating their research works to develop a software with the help machine learning algorithm which can help doctors to take decision regarding both prediction and diagnosing of heart disease. The main objective of this project is predicting the heart disease of a patient using machine learning algorithms. Comparative study of performance of machine learning algorithms is done.*

Key Words: Heart Disease; Prediction System; Machine Learning; Data Mining; Classification.

1.INRODUCTION

The successful treatment of a disease is always attributed by early and accurate diagnosis. Now a days doctors are adopting many scientific technologies and methodology for both identification and diagnosing not only common disease, but also many fatal diseases. We are living in an "information age" where we have large amounts of data being generated every day. This data can be used in accordance with various techniques of artificial intelligence and machine learning to effectively detect the presence of diseases in patients or the progression of certain diseases in patients. This process involves the effective recognition of information from huge amounts of data. This has been described as the process of "Knowledge Discovery from data", which can defined as the process of converting raw data into organized form, which consists of valuable information which can be used for decision making in many applications. Data mining techniques can be used to extract this information from raw data and convert it into suitable formats to be used. This information can then be used with machine learning algorithms to make predictions and classification.

These data mining and machine learning techniques can be used in the field of medicine for diagnosis of diseases. Working on heart disease patients is an application of data mining and machine learning techniques in this field.

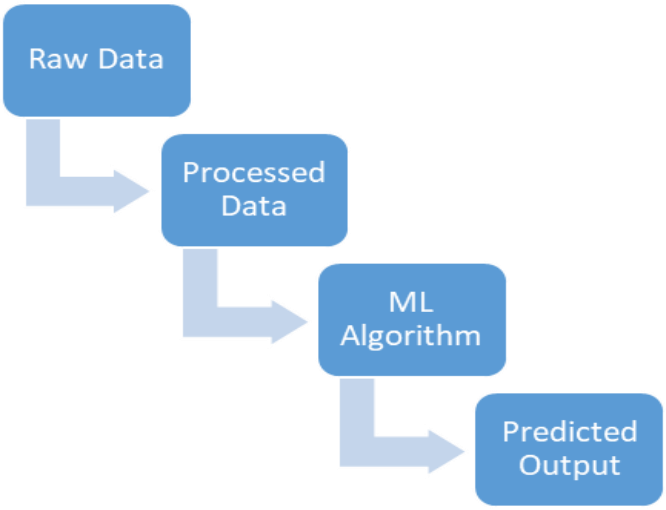


Fig 1: Block diagram of the prediction system

1.1 Types of Cardiovascular Diseases:

Heart diseases or cardiovascular diseases (CVD) are a class of diseases that involve the heart and blood vessels. Cardiovascular disease includes myocardial infarction (commonly known as a heart attack) and coronary artery diseases (CAD) like angina. Also coronary heart disease(CHD), in which a waxy substance called plaque develops inside the coronary arteries. Oxygen-rich blood is supplied by these artiries to heart muscle. When the plaque begins to build up in these arteries, the condition is called atherosclerosis. The development of plaque occurs over numerous years. As the time passes, this plaque can harden or rupture. The hardened plaque eventually narrows the coronary arteries. This causes reducuction of the flow of oxygen-rich blood to the heart. And if plaque ruptures, a blood clot can form on its surface. If the blood clot is large then it can completely block blood flowing through a coronary artery. As the time passes, the ruptured plaque also hardens and narrows the coronary arteries. If the stopped blood flow isn't treated on time, the section of heart muscle begins to die. Without timely treatment, a heart attack can lead to serious health problems and even death.

1.2 Prevalence of Cardiovascular Diseases:

An estimated 17.5 million deaths occur due to cardiovascular diseases all around the world. More than 75% of deaths occur due to cardiovascular diseases with-in the middle-

income and low-income countries. Also, 80% of the deaths that occur because of CVDs are result of stroke and heart attack . Every year there is increase of CVD patients in India. Currently, the amount of heart disease patients in India is more than 30 million. Over 2 lakh open heart surgeries are performed in India every year. A matter of growing concern is that the amount of patients requiring coronary interventions has been rising at 20% to 30% for the past few years.

1.3 Data Mining:

Data Mining has to be done on the raw dataset to extract useful information to be used in machine learning algorithm[8].

1.4 The Dataset:

The dataset has been imported from the University of California, Irvine repository for machine learning. The dataset is multivariate with total 75 attributes. The attributes consist of categorical, numeric, binary and continuous attributes. Out of these 75 attributes, 11 major attributes are considered for this problem. The total number of instances are 303. There are 5 labels which are 0 for no disease and 1-4 for the progression of disease. There are multiple datasets from different sources but Cleveland dataset is used here as it has a smaller number of missing values hence is more accurate. It consists of 0.2% missing values and is used for classification problem.[5]

1.4.1 Dataset Attributes Documentation:

The 11 main attribute which helps machine to learn are:-

1. Age : age in years

2. Thal : 3 = normal;
   6 = fixed defect;
   7 = reversable defect

3. Chest Pain : chest pain type
   - Value 1 - typical angina
   - Value 2 - atypical angina
   - Value 3 - non-anginal pain
   - Value 4 - asymptomatic

4. Number of major vessels : 0-3 colored by flourosopy

5. ST depression induced by exercise relative to rest.

6. Exercise induced angina : 1 = yes; 0 = no

7. Maximum heart rate achieved

8. Slope of the peak exercise ST segment :
   - Value 1 - Upsloping
   - Value 2 - Flat
   - Value 3 - Downsloping

9. Sex : sex (1 = male; 0 = female)

10. Rest ECG : resting electrocardiographic results
    - Value 0 - Normal
    - Value 1 - Having ST-T wave abnormality
    - Value 2 - Showing probable or definite left chamber hypertrophy

11. Diagnosis of heart disease : Value range between 0 to 4

1.4.2 Anomaly in Dataset:

During the analysis of the dataset it is found that the dataset is highly imbalanced. Apart from the label 0, the entries for the labels 1-4 are under sampled. This means that there is not enough data separately for these labels to effectively predict the presence of heart disease, which would result in low accuracy and precision. For example if a dataset consists of 100 instances out of which the tuples for label 1 are 98 and for the label 2 are only 2, we say that the data for label 2 is under sampled, so even if the classifier predicts all the inputs to be true it would be 98% accurate but actually it won't be able to classify correctly.

To deal with this problem, synthetic minority oversampling can be done by adding the instances of the under sampled labels, or the majority class data can be under sampled by removing its entries. With the kind of data available, under sampling and over sampling will not affect the accuracy of the classification much. Another way to deal with this problem is to define two labels, i.e. for positive or negative diagnosis. This balances the dataset and increases the accuracy and precision of classification.

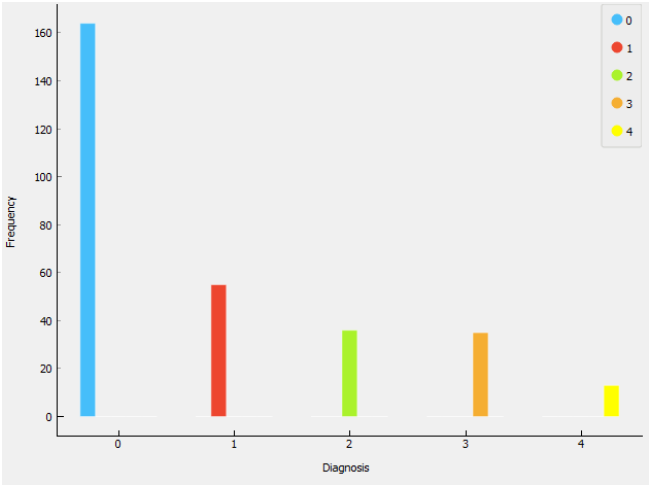


Fig 2: Classification of data according to Diagnosis

2. PROPOSED MECHANISM

In this project the training has been done using two algorithms namely, Support Vector Machine(SVM) and Random Forest[6].

2.1 Support Vector Machine:

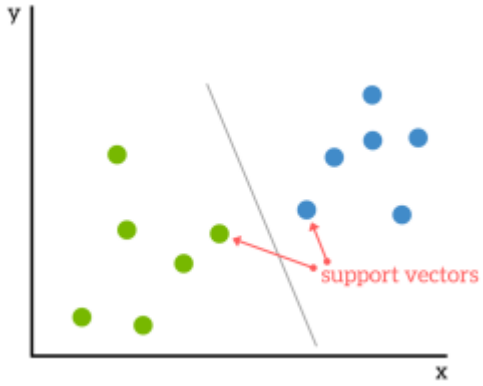
Support Vector Machines are based on the idea of graphically representing the data points in space and deriving a

geometrical shape which can segregate data points to classify data. SVM provides high performing algorithm without needing much optimization. It is one of the oldest and widely used machine learning algorithm used for classification.

It is a supervised machine learning algorithm, which can be used for mainly classification and often for regression purposes.

An SVM algorithm is based on finding a hyperplane, a geometrical figure which is defined in such a way that it can differentiate the data points in space. [1]An illustration of SVM is given below.

Support vectors are the data points in space nearest to the hyperplane, the points of a data set that would determine the position of the hyperplane. So support vectors can be considered most critical elements of a support vector machine.



What is a hyperplane?

A hyperplane can be explained as a geometrical subspace whose dimension is always one less than its surrounding ambient space. It can be a line in a two dimensional space, or a place in three dimensional space and so on.

Intuitively, the further the data points are from the hyperplane, the more confidently we can say that the classification is done properly. We therefore want our data points to be as far away from the hyperplane as possible, while still being on the correct side of it.

So when new testing data is added, whatever side of the hyperplane it lands will decide the class that it belongs to. This class has been assigned to it by us.

How do we find the right hyperplane?

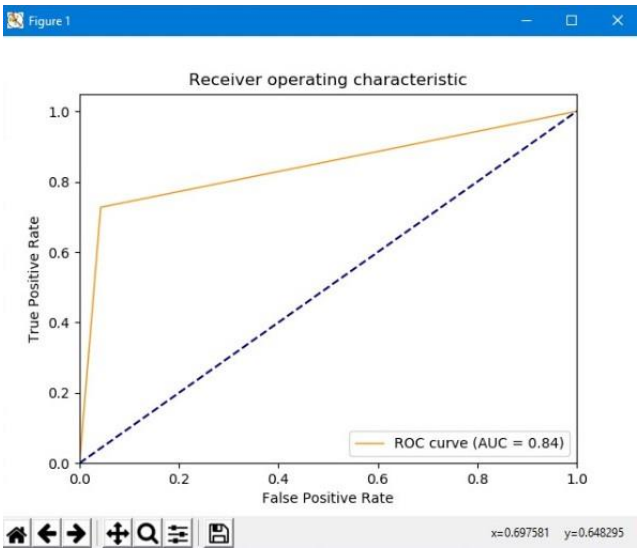
The distance from either set between the hyperplane and the nearest data point is called the margin. [2]The goal is to select a hyperplane with the greatest pos-siblemargin between the hyperplane and any point within the training set, making it more likely that new data will be properly classified.

But what happens when there is no clear hyperplane?

That's where it can get cumbersome. Data is rarely as simple and clean as ever. A dataset will often look more like the jum bled balls under which a linearly non-separable dataset is represented.

It is necessary to move away from a 2d view of the data to a 3d view in order to classify a dataset like the one above.[3]This is often referred to as kerneling. Now this classifier is a plane rather than a line. As we continue to increase the dimensions, the plane's dimensions also increase.

SVM is used for text classification tasks such as category assi gnment, spam detection and sentiment analysis.
It is also commonly used for challenges of image reco-gnition, performing especially well in aspect based recognition and classification based on color.
In many areas of handwritten digital recognition, such as pos tal automation services, SVM also plays a vital role.



2.2 Random forest:

What is random forest?

A Random Forest is a classifier consisting of collection of tree-structured classifiers where independent random vectors are distributed identically and each tree cast a unit vote for the most popular class at input x.A random vector is generated which is independent of the past random vectors with same distribution and a tree is generated by using the training test[6]. For random forests, an upper bound is derived to obtain the generalization error in terms of two parameters that are given below:

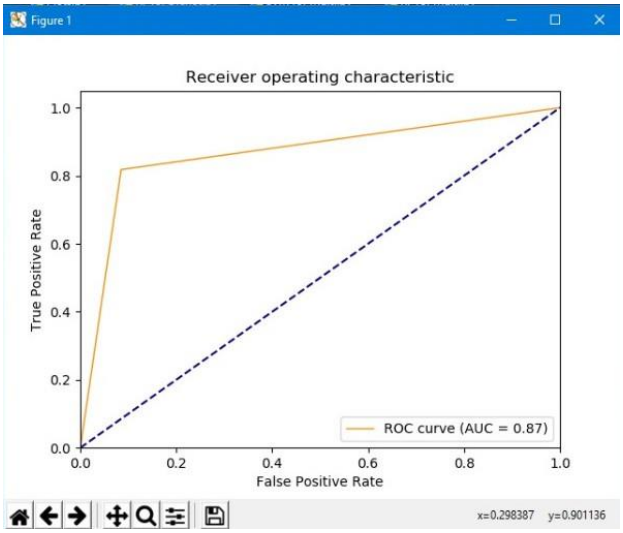
- The accuracy of individual classifiers
- The dependency between the individual classifiers
- The generalization of error for random forest includes two segments. These segments are defined below:
- The strength of the individual classifiers in the forest.
- The correlation between them in terms of raw margin function

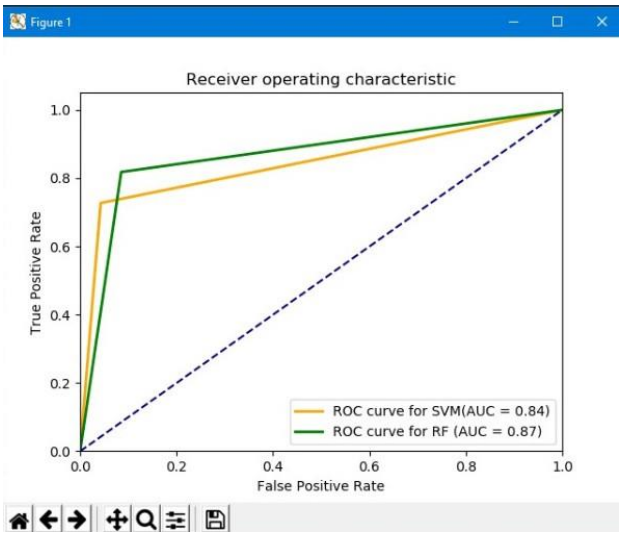
How random forests work ?

- A different subset of the training data are selected (~2/3), with replacement, to train each tree[4].
- Remaining training data (OOB-out of box) are used toestimate error and variable importance
- Class assignment is made by the number of votes from all of the trees and for regression the average of the results is used

Advantages of random forests :

- No need for pruning trees
- Accuracy and variable importance generated automatically
- Overfiting is not a problem
- Not very sensitive to outliers in training data
- Easy to set parameters





2.3 Graphical User Interface:-

The gui takes the data of the attributes which helps in predicting the heart disease. The gui is implemented using both the algorithms which enables the analyzer to calculate the difference between both the predictions.

3. RESULT

The accuracy obtain for the two mechanisms totally depends on the data set used, also the prediction depends on two types i)detecting the presence of heart attack ii) how much severe is the condition of the heart attack.
The accuracy for detecting the presence of heart attack is for support vector machine is 84.61% and for random forest is 86.81% and how much severe is the heart attack is for svm is 60.44% and for random forest is 59.34%.

4. CONCLUSION

Heart diseases when diagnosed early can be managed by various ways. By using the above approach we can predict the presence oh heart disease using the various symptoms of a patient. The two classifiers used give the most accurate predictions in this case, i.e. support vector machine and random forest classifier. Due to the unavailability of data in abundance we cannot predict the heart diseases of different kinds accurately, but the diagnosis of heat disease can be done with a fair accuracy of about 80 – 85 %. More accurate systems can be developed in future for diagnosis of diseases, when enough data is available.

5. REFERENCES

[1] Shashikant Ghumbre, Chetan Patil, and Ashok Ghatol," Heart Disease Diagnosis using Support Vector Machine", International Conference on Computer Science and Information Technology (ICCSIT'2011) Pattaya Dec. 2011

[2] Prashasti Kanikar, Disha Rajeshkumar Shah ," Prediction of Cardiovascular Diseases using Support Vector Machine and Bayesian Classification", International Journal of Computer Applications (0975 – 8887) Volume 156 – No 2, December 2016

[3] Mythili T., Dev Mukherji, Nikita Padalia, and Abhiram Naidu," A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL)", International Journal of Computer Applications (0975 – 8887) Volume 68– No.16, April 2013

[4] Prof. Priya R. Patil, Prof. S. A. Kinariwala," Automated Diagnosis of Heart Disease using Random Forest Algorithm", International Journal of Advance Research, Ideas and Innovations in Technology

[5] Jaymin Patel, Prof.TejalUpadhyay, Dr. Samir Patel," Heart Disease Prediction Using Machine learning and Data Mining Technique", International Journal of computer science and communication.

[6] T. Marikani, K. Shyamala," Prediction of Heart Disease using Supervised Learning Algorithms", International Journal of Computer Applications (0975 – 8887) Volume 165 – No.5, May 2017

[7] S.Nandhini, Monojit Debnath , Anurag Sharma , Pushkar, " Heart Disease Prediction using Machine Learning", International Journal of Recent Engineering Research and Development (IJRERD) ISSN: 2455-8761 www.ijrerd.com || Volume 03 – Issue 10 || October 2018

[8] Maqsud S. Kukasvadiya, Dr.Nidhi H. Divecha," Analysis of Data Using Data Mining tool Orange", 2017 IJEDR | Volume 5, Issue 2 | ISSN: 2321-9939