International Journal of Technical Research and Applications e-ISSN: 2320-8163,
www.ijtra.com Volume 6, Issue 3 (MAY-JUNE 2018), PP. 09-13

HEART DISEASE PREDICTION USING DATA MINING ALGORITHMS

Abhishek Verule^1, Prof. Shalini L^2,
SCOPE, VIT University, Vellore
[1]VIT University, Vellore
[2]VIT University, Vellore
Vellore, India

Abstract— **The age we are living is an 'information age'. Every day, terabytes of data are produced. Data mining is the practice of examining large pre-existing databases in order to generate new information. A huge amount of data is produced in the healthcare industry every day. However, most of this data is not used efficiently. Data mining techniques and machine learning algorithms play a vital role here. Prediction and diagnosing of heart disease become a challenging factor faced by doctors and hospitals both in India and abroad. This paper summarizes some of the current exploration on heart disease prediction using data mining algorithms, analyze and compare them to conclude which technique is more effective and efficient.**

Index Terms— **Logistic regression, Decision tree, Support Vector Machine (SVM).**

I. Introduction

The heart is an important muscular organ in most animals. Heart plays a crucial role in the circulatory system. The circulatory system functions to transport blood and oxygen from the lungs to the various tissues of the body. The heart pumps the blood throughout the body. If the heart does not function properly, it will lead to serious health conditions including death.

A. Types of Cardiovascular diseases –

Heart diseases or cardiovascular diseases (CVD) are class of diseases that involve the heart and blood vessels. CAD refers to the narrowing of the coronary arteries, the blood vessels that supply oxygen and blood to the heart. It is also known as coronary heart disease (CHD). CHD starts with injury or damage to the inner layer of a coronary artery. This damage causes fatty plaque deposits to build up at the site of the injury. These deposits consist of cholesterol and other cellular waste products. The accumulation is called atherosclerosis. This clump can block the artery, reducing or blocking blood flow, and leading to a heart attack. It is a major cause of illness and death. If the stopped blood flow isn't restored quickly, the section of heart muscle begins to die. Without quick treatment, a heart attack can lead to serious health problems and even death. Heart attack is a common cause of death worldwide. Some of the common symptoms of heart attack are as follows:-

1) *Chest pain*

People describe it as a squeezing, pressure, heaviness, tightening, burning, or aching across the chest. It usually starts behind the breastbone. The pain often spreads to the neck, jaw, arms, shoulders, throat, back, or even the teeth. It is the most common symptom of heart attack.

2) *Shortness of breath (dyspnea)*

CHD can lead to shortness of breath. If the heart and other organs are getting too little oxygen, the patient may start panting. Any exertion may be very tiring. Nausea, Indigestion, Heartburn and Stomach Pain.

3) *Pain in the Arms*

The pain usually moves from the chest towards the arms, especially on the left side.

4) *Fatigue*

Simple routines which begin to set a feeling of tiredness should not be ignored.

5) *Feeling Dizzy and Light Headed*
6) *The loss of balance.*

Some other cardiovascular diseases which are quite common are heart failure, stroke, Cardiomyopathy, hypertensive heart disease, peripheral artery disease, Congenital heart disease, Venous thrombosis, Valvular heart disease, Aortic aneurysms, rheumatic heart disease and heart arrhythmia. Certain abnormalities in the functioning of the circulatory system may be the cause of the development of heart disease. These abnormalities or heart disease itself may develop due to caertain lifestyle choices such as being physically inactive, excess alcohol and tobacco intake, an unhealthy diet with excess sugar and salt and others. Early detection of heart disease leads to proper treatment and helps keep the disease under control.

B. Prevalence of Cardiovascular Diseases

An estimated 17.5 million deaths occur due to cardiovascular diseases worldwide. More than 75% deaths due to cardiovascular diseases occur in the middle-income and lowincome countries. Also, 80% of the deaths that occur due to CVDs are because of stroke and heart attack. India too has a growing number of CVD patients added every year. Currently, the number of heart disease patients in India is more than 30 million. Over two lakh open heart surgeries are performed in India each year. A matter of growing concern is that the number of patients requiring coronary interventions has been rising at 20% to 30% for the past few years.

The rest of the paper is organized as follows. Section 2 describes the data mining algorithms used for heart disease prediction. Section 3 is a literature survey on papers referred. Section 4 comprises of the results of the experiment. Finally, Section 5 concludes the paper along with future scope.

II. METHODOLOGY

Many data mining algorithms have been formulated based on the research in data mining. When applied directly on a dataset, models can be created and significant conclusions or inferences can be made based on the dataset. The data mining algorithms compared in this paper are-

A. Decision Tree

A decision tree is a tree where each node represents an attribute, each branch represents a decision and each leaf represents an outcome. The idea is to create a tree like this for the entire data and process a single outcome at every leaf (or minimize the error in every leaf). A Decision tree is a graph of decisions and their possible outcomes including chance-related outcomes. It is visualized as a tree-like model. It is one of the ways to display an algorithm. Decision trees are commonly used in machine learning, specifically in decision analysis to aid the identification of a strategy that will most likely reach the required outcome. It is also a popular tool in operations research. The mappings from the root node to the leaf nodes (one by one) allow us to transform a decision tree to a set of rules. Finally, by following these set of rules, useful conclusions can be achieved.

B. Support Vector Machine (SVM)

SVM is a supervised machine learning algorithm. The algorithm works for both classification and regression challenges. However, it is common in classification problems. In this algorithm, each data item is plotted as a point in n-dimensional space (where n is number of features present) with the value of each feature being the value of a particular coordinate. Then, a hyper-plane is found that separates the two classes. Support Vectors are simply the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes (hyper-plane/ line).

Support vector machine performs a similar task like C4.5 except that it doesn’t use Decision trees at all. Support vector machine attempts to maximize the margin (distance between the hyper plane and the two closest data points from each respective class) to decrease any chance of misclassification. Some popular implementations of support vector machine are scikit-learn, MATLAB and LIBSVM.

C. Logistic Regression

David Cox, who was a statistician, introduced Logistic regression in 1958. The binary logistic model is used to measure the probability of a binary response based on one or more independent features. As a risk factor is present, it increases the odds of a given outcome by a specific factor. The model is not a classifier but a direct probability model.

In statistics, logistic regression is a regression model where the dependent variable is categorical. A categorical variable can take on one of a number of possible values that are limited and generally fixed, assigning each individual or other unit of observation to a particular group on the basis of some qualities. In the case of a binary dependent variable, the output can take only two values, "0" and "1", which represent outcomes such as failed/not failed, win/lose, dead/not dead or sick/not sick. Cases where the dependent variable has more than two result values can be resolved in multinomial logistic regression, or, if the multiple categories are ordered, in ordinal logistic regression. Logistic regression is an example of a qualitative response/discrete choice model.

Logistic regression can help us to find a function between dependent variable and independent variables. The decision boundary is linear as it is a linear classifier. Classification problems can be solved by using Logistic regression.

III. LITERATURE SURVEY

AH Chen et al. presented a heart disease prediction system that can help doctors in determining heart disease status based on the clinical data of patients. 13 important clinical features such as age, sex, chest pain type were chosen. The heart disease was classified using an artificial neural network (ANN) based on these clinical attributes. Data was collected from the UCI repository. The ANN model contained three layers i.e. the input layer, the hidden layer and the output layer having 13 neurons, 6 neurons and 2 neurons respectively. Learning Vector Quantization (LVQ) was used in this study. LVQ is a special case of an ANN that applies a prototype-based supervised classification algorithm. The heart disease classification and prediction was implemented using C programming language and trained by an artificial neural network. The system was developed in C and C# environment. The accuracy of the proposed method for prediction is near to 80%.

Mrudula Gudadhe et al. presented a decision support system for heart disease classification. Support vector machine (SVM) and artificial neural network (ANN) were the two main methods used in this system. A multilayer perceptron neural network (MLPNN) with three layers was employed to develop a decision support system for the diagnosis of heart disease. This multilayer perceptron neural network was trained by back-propagation algorithm which is a computationally an effectual method. Results showed that a MLPNN with back-propagation technique can be successfully used for diagnosing heart disease.

K. Sudhakar et al. studied heart disease prediction using data mining. The data generated by the healthcare industry is huge and “information rich”. As such, it cannot be interpreted manually. Data mining can be effectively used to predict diseases from these datasets. In this paper, different data mining techniques are analyzed on heart disease database. In this paper, techniques such as Decision tree, Naïve Bayes and neural network were applied. Associative classification is a technique which combines associative rule mining and classification to a model for prediction to give maximum accuracy. In conclusion, this paper analyzes and compares how different classification algorithms work on a heart disease database.

Kamal Kant et al. proposed a prototype of heart disease prediction using data mining techniques, namely Naïve Bayes. Naïve Bayes is a statistical classifier which assigns no dependency between the attributes. The posterior probability needs to be maximized for determining the class. Here, Naïve Bayes classifier also performs well. Naive Bayes was observed as the most effective model for disease prediction before neural networks and Decision trees.

Shadab Adam Pattekari et al. developed a prototype of Heart Disease Prediction System using Naive Bayes, Decision trees and neural networks. It is implemented in a web application. In this system, user answers some predefined questions. Then it retrieves hidden data from the stored database and compares the user’s values with trained dataset. Hidden knowledge associated with heart diseases can be found and extracted by this system from a heart disease database. It can answer the complex queries for diagnosing a disease. Naive Bayes classification algorithm was applied on a set of 15 attributes to find the chances of heart disease.

Boshra Baharami et al. compared different classification techniques such as Decision tree k-Nearest Neighbors (k-NN), Naive Bayes(NB) and SMO(training SVM). On the dataset feature selection technique (gain ratio evaluation technique) is used to extract the important features. WEKA software is used for implementing the classification algorithms. 10 fold cross-validation technique is used to test the mining techniques. J48 shows the highest accuracy of 83.732%.

Nidhi Bhatla et al. analysed various data mining techniques introduced in recent years to predict heart diseases. The results showed that neural networks with 15 attributes perfomed better than all other data mining techniques and the Decision tree also showed good accuracy with the help of genetic algorithm. Apart from the common attributes, this research work was done by using two more attributes i.e. obesity and smoking for productive diagnosis. Genetic algorithm was applied which uses natural evolution methodology. It continues generation until it evolves a population P where every rule in P satisfied the fitness threshold, starting from null. Decision tree gave an accuracy of 99.62% by using 15 attributes. Also, when combined with genetic algorithm with 6 attributes, Decision tree gave an efficiency of 99.2%.

IV. EXPERIMENTAL RESULTS

Comparison in terms of accuracy, true positive rate and false positive rate was conducted. Also, models were evaluated in terms of AUC measure.

Since, we are dealing with disease investigation, we want to reduce cases when we conclude that a person is not sick while he/she is sick and so the best model is selected using true positive rate, which is a fraction of TP/actual yes.

The data was collected from four sources:

1) Cleveland Clinic Foundation (cleveland.data)
2) Hungarian Institute of Cardiology, Budapest (hungarian.data)
3) V.A. Medical Center, Long Beach, CA (long-beach-va.data)
4) University Hospital, Zurich, Switzerland (switzerland.data)

We split the data into training and testing sets (75% vs 25%). We used training set to create models and testing set for assessing predictive power of our models.

AUC stands for "Area under the Receiver Operating Characteristic curve (ROC Curve)”. That is, AUC measures the total two-dimensional area below the total ROC curve from (0,0) to (1,1). AUC provides an aggregate measure of performance across all possible classification thresholds. One way of understanding AUC is as the probability that the model ranks a random positive example more highly than a random negative example.

Sl. No.		Name of Algorithm	AUC Metric
1	i.	Decision Tree-Gini	0.8724564
	ii.	Decision Tree-Entropy	0.8710029
2		Logistic Regression	0.8968023
3		Support Vector Machine	0.9382267

A. Logistic Regression

Logistic regression model gives an accuracy of 83% and a true positive rate of 0.81 which is not good enough. This means that out of 100 patients, 19 patients who are actually sick will be diagnosed as not sick. This is a big risk and hence the model cannot be assumed as the best one. Also, the logistic model gives an AUC metric of 0.8968023 which is the second best and is impressive.

B. Decision tree

While performing the Decision Tree algorithm, trees using both Gini and Entropy were formed. They can be seen in figures 1 and 2.

Based on the table. we can notice for the data, the model that uses entropy measure gives higher accuracy as well as false and true positive rates. If we compare with logistic model, model based on entropy gives us higher true positive rate, but lower false positive rate. AUC metric of both Gini based and

Entopy based Decision tree give 0.87 with the Gini based tree's metric just exceeding that of the entropy based tree.

TABLE I
ACCURACY AND POSITIVE RATES FOR MODELS

Sl. No.	Name of Algorithm	Accuracy (%)	False Positive	True Positive
1	Decision Tree-Gini	0.76	0.26	0.78
	Decision Tree-Entropy	0.83	0.23	0.91
2	Logistic Regression	0.83	0.16	0.81
3	Support Vector Machine	0.85	0.14	0.84

C. Support Vector Machine

The SVM model gives the best accuracy percentage and the lowest false positive rate. Also, it is only the second best in terms of true positive rate; only behind entropy based decision tree. SVM also gives the best AUC metric amongst the selected models. But, since our aim is to increase the true positive rate, it is the second best model behind entropy-based decision tree.

AUC METRIC FOR THE MODELS

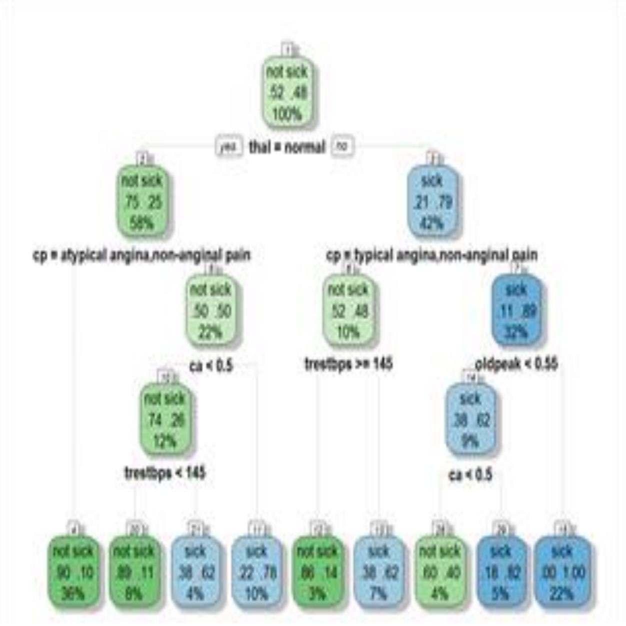

Fig.1. Tree based on Gini Index

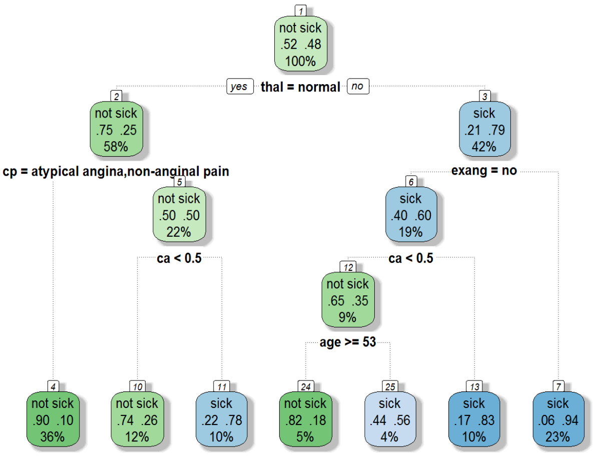

Fig.2. Tree based on entropy

V. CONCLUSION

Heart disease is one of the major causes of death in the world. So, the early detection of Heart disease is needed to reduce life losses. In this paper we have compared three different models for data mining namely, Logistic Regression, Decision tree and Support Vector Machine. On comparing the models based on their true positive rate, our study reveals that Entropy based decision tree model gives the best results at least when compared to the other methods. This work can further be enhanced in near future.

A smart system may be developed in the future that can lead to selection of proper treatment methods for a positively diagnosed patient. A lot of work has been done already in making models that can predict if a patient is likely to develop heart disease or not. There are several treatment methods for a patient once diagnosed with a particular form of heart disease. Data mining can be of very good help in deciding the line of treatment to be followed by extracting knowledge from certain databases.

REFERENCES

[1] Animesh Hazra, 2Subrata Kumar Mandal, 3Amit Gupta, 4Arkomita Mukherjee and 5Asmita Mukherjee. Heart Disease Diagnosisand Prediction Using Machine Learning and Data Mining Techniques : A Review

[2] Purushottama,c* , Prof. (Dr.) Kanak Saxenab, Richa Sharmac. Efficient Heart Disease Prediction System

[3] Ilayaraja M*, Meyyappan T. Efficient Data Mining Method to Predict the Risk of Heart Diseases through Frequent Itemsets

[4] Ramandeep Kaur, Er. Prabhsharn Kaur. A Review - Heart Disease Forecasting Pattern using Various Data Mining Techniques.

[5] Shadab Adam Pattekari,and Asma Parveen, 2012,"Prediction System for Heart Disease using Naive Bayes", International Journal of Advanced Computer and Mathematical Sciences, ISSN: 2230-9624,Vol. 3, Issue 3.

[6] AH Chen, SY Huang, PS Hong, CH Cheng, and EJ Lin,2011, "HDPS: Heart Disease Prediction System",Computing in Cardiology, ISSN: 0276-6574.

[7] Mrudula Gudadhe, Kapil Wankhade, and Snehlata Dongre, Sept 2010,"Decision Support System for Heart Disease Based on Support Vector Machine and Artificial Neural Network",International Conference on Computer and Communication Technology.

[8] K.Sudhakar, and Dr. M. Manimekalai, January 2014, "Study of Heart Disease Prediction using Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, Issue 1.

[9] Kamal Kant, and Dr. Kanwal Garg,2014, "Review of Heart Disease Prediction using Data Mining Classifications",International Journal for Scientific Research & Development(IJSRD), Vol. 2, Issue 04, ISSN (online): 2321-0613,

[10] Boshra Bahrami, and Mirsaeid Hosseini Shirvani,February 2015,"Prediction and Diagnosis of Heart Disease by Data Mining Techniques",Journal of Multidisciplinary Engineering Science and Technology(JMEST)

[11] Nidhi Bhatla, and Kiran Jyoti, Oct. 2012,"An Analysis of Heart Disease Prediction using Different Data Mining Techniques", International Journal of Engineering Research & Technology (IJERT)

[12] Indira S. Fal Dessai,2013,"Intelligent Heart Disease Prediction System Using Probabilistic Neural Network", International Journal on Advanced Computer Theory and Engineering (IJACTE)

[13] Mai Shouman, Tim Turner, and Rob Stocker, June 2012,"Applying k-Nearest Neighbors in Diagnosing HeartDiseasePatients",International Journal of Information and Education Technology, Vol. 2, No. 3.

[14] Serdar AYDIN, Meysam Ahanpanjeh,and Sogol Mohabbatiyan,February 2016, "Comparison And Evaluation of Data Mining Techniques in the Diagnosis of Heart Disease",International Journal on Computational Science & Applications (IJCSA), Vol. 6,No.1.

[15] G. Purusothaman, and P. Krishnakumari, June 2015,"A Survey of Data Mining Techniques on Risk Prediction: Heart Disease", Indian Journal of Science and Technology.

[16] Deepali Chandna, 2014,"Diagnosis of Heart Disease Using Data Mining Algorithm", International Journal of Computer Science and Information Technologies (IJCSI).

[17] Dhanashree S. Medhekar, Mayur P. Bote, and Shruti D. Deshmukh, March 2013,"Heart Disease Prediction System Using Naive Bayes", International Journal of EnhancedResearch in Science Technology & Engineering.

[18] Noura Ajam, 2015, "Heart Diseases Diagnoses Using Artificial Neural Network", Network And Complex Systems.

[19] Vikas Chaurasia, and Saurabh Pal, 2013, "Early Prediction of Heart Diseases Using Data Mining Techniques", Caribbean Journal of Science and Technology.

[20] Jaymin Patel, Prof. Tejal Upadhyay, and Dr. Samir Patel, Sep 2015-Mar 2016, "Heart Disease Prediction using Machine Learning and Data Mining Technique".

[21] Serdar AYDIN, Meysam Ahanpanjeh,and Sogol Mohabbatiyan,February 2016, "Comparison And Evaluation of Data Mining Techniques in the Diagnosis of Heart Disease",International Journal on Computational Science & Applications (IJCSA).