ISSN XXXX XXXX © 2017 IJESC

Research Article Volume 7 Issue No.12

Heart Disease Prediction using Data Mining Techniques

Nallakaruppan.M.K[1], N. Deepa[2], Pranava Kumar R[3], Moin Ahmed[4], Tanuja[5], Ajith Kumar[6]
Assistant Professor[1, 2, 3], M. Tech Student[4, 5, 6]
Department of School of Information Technology & Engineering
Vellore Institute of Technology, Vellore, Tamilnadu, India

Abstract:
Heart Diseases refer to an abnormality in the functioning of the heart. These have become one of the major causes of death in our country. They occur all of a sudden and it becomes difficult to diagnose as we may have little time. Within this time the disease needs to be treated with utmost care and precision. Also the treatment of disease is risky and expensive. The probability of loosing life due to this attack is very high. Huge amount of data is available in the medical industry regarding heart diseases. There is lack of effective analysis tool to discover hidden relationships and trends in data. Here we are going to find out the accuracy and effective relations among them using data mining techniques such as clustering algorithms like k-means and other classification techniques. Thus the comparison of everything by research and publication of analysis will be done.

Keywords: KNN (K-nearest neighbor),Weka(Waikato Environment for Knowledge Analysis), csv(comma separated values)

I. INTRODUCTION

Heart diseases are classified mainly into two types namely congenital and acquired .Babies born with a heart disease are considered to have congenital heart disease and some get it at later stage.

It is called as acquired heart disease. Majority of the people get acquired heart diseases. It is very usual that people with heart disease to have symptoms.

They may have the symptoms like chest pain, no proper distribution of blood through lungs causing difficulty in breathing, changing of skin tone into blue, swelling of legs or feet much because of blood backing up to lower part of the body from heart.

There is also a probability for a person with heart disease not showing any symptoms. In such a situation there is an urgency to predict the disease at an early stage as soon as possible. Data mining means extracting useful information from a large database.

Data mining techniques are used to discover the interesting patterns in least amount of time. In the recent years data mining has become very important in the health care in prediction of diseases because it is useful in detecting the important and undiscovered data patterns in the medical data more efficiently.

The data extracted must be carefully analyzed and executed much accurately and precisely in medical diagnosis.

As heart diseases are predominant now a days the data regarding heart disease is collected and diagnosis of heart disease using various data mining techniques is done. Various existing data mining techniques are executed and compared to discover the most accurate and precise technique in predicting the heart disease.

The main aim of this paper is to identify the most appropriate data mining technique to predict the heart disease at an early stage by analyzing different predictive data mining techniques.

II. METHODOLOGY

The methodology we use in this paper mainly is doing a survey on heart diseases.

We collect the data with attributes pertaining to heart disease prediction.

Then using weka different classification and clustering mechanisms are tried for the heart disease dataset and the best among the techniques are chosen. The dataset is also tested in Rstudio for accurate results,

The results obtained from various classifications and clustering algorithms is compared and the analysis of the results is done using R programming.

The work is mainly focused on discovering the attributes that mainly lead to disease, data mining technique that is useful in predicting the heart disease most accurately and precisely at an early stage.

III. DATASET DESCRIPTION

The dataset we used in this paper is Cleveland dataset containing 303 instances. The dataset that is recently updated in the data world library is taken.

This dataset contains total 76 attributes out of which only 14 attributes are used in this paper for prediction of heart disease

Table.1.Dataset description

Attribute name	Attribute description	Attribute type
Age	Patient age in years	Numerical
Sex	Sex (1-male;0=female)	Numerical
Cp	Chest pain type (1:typical angia,2:atypical angina,3:non-anginal pain,4:asymptomatic)	Numerical
trestbps	Resting blood pressure	Numerical
Chol	Serum cholesterol measured in mg/dl	Numerical
Fbs	Fasting blood sugar(>120 mg/dl 1:true,0:false)	Numerical
restecg	Resting electrographic results[0-2]	Numerical l
thalach	Heart rate maximum achieved	Numerical
exang	Induced angina due to exercise(1:yes,0:no)	Numerical
oldpeak	ST depression induced due to exercise relative to rest	Numerical
Slope	Slope of peak ST exercise segment(1:upsloping,2:flat,doensloping)	Numerical
Ca	No. of blood vessels colored by fluoroscopy(0-3)	Numerical
Thal	3:normal,6:fixed defect,7:irreversible defect	Numerical
Num	Diagnosis of heart disease(prediction attribute)	Numerical

IV. CLASSIFICATION ALGORITHMS

k-star:

K-star is a lazy learning algorithm of weka. It is used to find the k shortest paths in a directed graph between the chosen pair of vertices. This algorithm doesn't necessarily require the weighted graph to be stored in the main memory and hence the graph need not be available explicitly. According to necessisity portions of graph will be generated star uses a distance function based on entropy. Every test instance class of it is based on class of similar training instances.

IBk:

It is a also a lazy classification algorithm that implements k-nearest neighbors algorithm. Distance weighting can also be done using this algorithm. Based on cross validation we can choose the k value.

Random Tree

Basically it is a Classification Algorithm/Process; Random Committee belongs to trees classifier. Random tree is a class for constructing a tree that chooses K randomly chosen attributes at each node. Random Tree is a Supervised Classifier. This algorithm can deal with both classification and regression problems. Random trees are a group of tree predictors that is called forest. It performs no trim. Also has an option to allow estimation of class probabilities based on hold-out set.

Random Forest

Basically it is a Classification Algorithm/ Process, Random Committee belongs to trees classifier. Random Forest is a class for constructing a forest of random trees. It is a general technique of random decision forests that are group learning technique for classification.

Random Committee

Basically it is a Classification Algorithm/ Process, Random Committee belongs to meta classifier. Random Committee is a Class for building an ensemble of arbitrary classifiers. Each base classifiers are built using a various number seed based on the same data. The final process/Prediction is a straight average of the predictions generated by the individual base classifiers

V. CLUSTERING ALGORITHMS:

Filtered Clustering:

It is a class for running an arbitrary clusterer on data that has been passed through an arbitrary filter. Like the clusterer the structure of the filter is based exclusively on the training data and test instances will be processed by the filter without changing their structure

Make density based clustering:

Make density based cluster is an clustering algorithm that supports the interface of requested no. of clusters only when the wrapped class supports it. It is a class that wraps the clustered to return density and distribution. This algorithm fits the discrete and normal distributions within each cluster generated by wrapped clusterer.

Simple k-means:

K-means algorithm provides a simplest way to classify the given dataset using k no. of clusters. We define k centroids, one for each cluster. In this algorithm we choose k no. of clusters (k fixed) ,find the centroids and distance between the objects to centroids, then we group them based o minimum distance loop is generated and iterated until there are no more changes to be done to change the centroids position after every loop. The main aim of this algorithm is to minimize the objective function

VI. RESULT ANALYSIS AND DISCUSSION:

Using WEKA:

Here in result and analysis we are going to tabulate and compute all the search results of classification and clustering. In this paper of predicting heart diseases we are going to compare 14 of the 76 attributes. Of these 76 attributes many are deemed to be not used and only 14 of them are useful for analysis. Initially the dataset taken is preprocessed in weka to remove the missing values and the data is normalized. After this preprocessing the obtained cleaned data is subjected to various classifications and clustering techniques .Then the statistics like mean absolute error, correlation coefficient, and relative absolute error are checked using these statistics we are going to determine the attribute that is acting as a deciding factor in predicting the heart disease. We have selected various attributes like chol, thal, num, restecg, fbs from the acquired dataset and tested each attribute against all the algorithms applicable for the heart disease dataset Of all the classifications done based on these attributes it is found that algorithms like IBK, Kstar, Random Committee, Random Tree, Random forest are efficient in predicting heart diseases. It was found that Mean Absolute Error (MAE) and Relative Absolute Error (RAE) of IBK to be very low like 0.066 and 0.65%.The correlation coefficient value were close to 0.997. In similar way the values of other classification algorithms that work well to predict the heart disease are listed below in the table. All these operations have been performed by taking num attribute as a class. That means that prediction of heart diseases is measured with respective to num value. Num denotes the diagnosis of heart disease. If value is 0 then diameter narrowing is <50% or value is >0 then diameter narrowing becomes >50%. Here if value is >0 it is divided as 50_1,50_2,50_3,50_4. With the help of this nominal attribute we are able to predict which attribute is likely more responsible for heart diseases.

Table.2.Classification technique results using weka

S.No	Classification Technique	Correlation Coefficient	Mean Absolute Error	Relative Absolute Error
1	**IBK**	0.997	0.066	0.65%
2	**K-Star**	1	0.0002	0.015 %
3	**Random Forest**	0.9763	0.239	23.58 %
4	**Random Tree**	0.9947	0.203	1.999 %
5.	**Random Committee**	0.999	0.0038	0.373 %

Other than classify us also has other technique called clustering. Here the techniques used majorly were Filtered Clusterer, Make density based cluster, Simple KMeans. Attributes such as age and sex can be ignored, after that various techniques were applied and results were compiled in different cluster modes. Of all the above mentioned three were successful in grouping mostly correct instances rather than incorrect instances. In Simple k Means at first on using use training dataset clustered sum of squared errors is obtained. Later we change the number of iterations in order to reduce the clustered sum of squared errors. It provides mean, number of clusters and clustered instances. Later when we use classes to clusters evaluation as cluster mode then in addition to the above we also obtain classes to clusters and incorrectly clustered instances. they help to predict the efficiency of dataset. Also we have Make Density Based clustering where get mean and standard deviation for numeric attribute whereas counts for nominal attributes. Later on applying classes to clusters evaluation we also get Log likelihood, class attribute, classes to clusters, incorrectly clustered instances.

Table.3.MakeDensityBasedClustering Analysis

Numeric attribute	Mean	Standard deviation	Nominal attribute	Count
age	57.2239	8.1062	sex	28 108(136)
trestbps	134.306	19.0331	cp	12 110 12 4(138)
thalach	253.667	55.9051	fbs	22 114 (136)
chol	137.7463	20.787	restecg	91 42 4 (137)
oldpeak	1.6187	1.2826	exang	47 89 (136)
Ca	0.9826	1.0151	Slope thal	24 99 14(137) 16 33 88(137)

Later Filtered Clustering is performed. Here obtain clustered centroids and clustered instances on full training set are obtained. Later on applying classes to clusters evaluation we get clustered instances, class attribute, classes to clusters, and incorrectly clustered instances in model and evaluation on training set. Also we obtain graphs which help to visualize and enhance the available data to predict heart diseases.

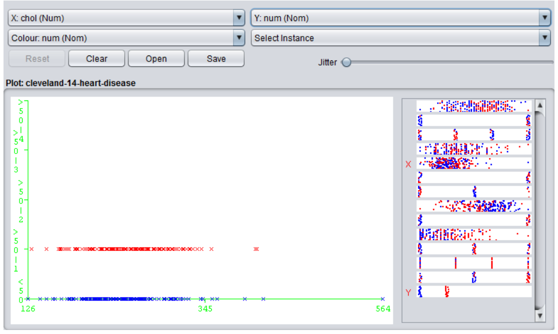

X: chol (Num)
Y: num (Nom)
Colour: num (Nom)
Select Instance
Plot: cleveland-14-heart-disease

Figure1.1.Filteredclusterer plot1

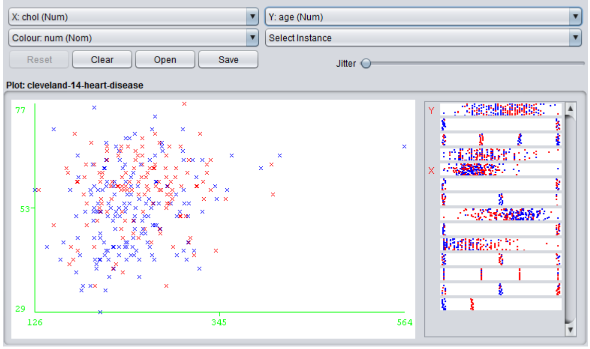

X: chol (Num)
Y: age (Num)
Colour: num (Nom)
Select Instance
Plot: cleveland-14-heart-disease

Figure.1.2.Filteredclusterer result plot2

VII. USING R:

K-NN:

Based on cross validation here k value is chosen. Of the 303 instances we have first 200 are divided into train data and the remaining instances into test data.. Later knn function is applied with train data to test data to find out number of correct instances taken. Here out of 103 instances in test data 21 instances in F and 53 instances in M are proven to be correct. Therefore we could achieve about 75% accurateness using this algorithm.

Table.4.K-NN values

Wdbc_pred	F	M
F	21	10
M	19	53

Random Forest:

It is done by grouping of various random trees. Here at first graph is plotted using box plot and mosaic plot for reference. On analyzing we found out that num and chol are the attributes which mainly lead to accurate data. Later the dataset is partitioned into train and test data. Random Forest function “rt” is applied to it. Using predict function we can find the random value of train data. Later tabulate table for test data with train data. Finally we get a random tree displaying the analyzed data:

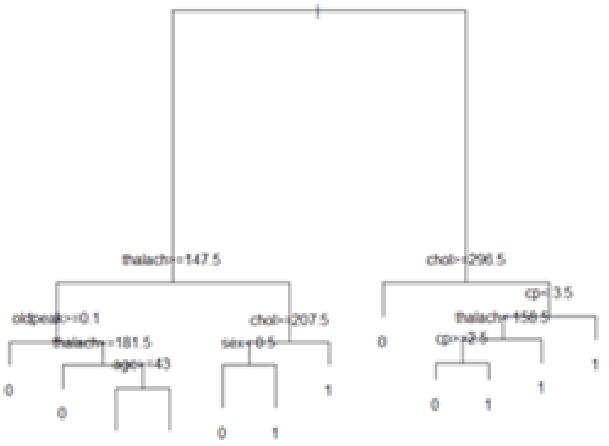


Figure.2.Output of Random Forest

K-means Clustering:

Here we are going to find the attributes which predict heart disease value with good accuracy. We are going to perform it by the using number of clusters. Here we are going to give 5 clustered values for K-means function. The results we obtain are means, vectors, sum of squares. By plotting the graph we obtain which attributes suit it in a best way. By analyzing the attributes having num and chol. It resembles like: Here in Fig 1.1 we have a plot with chol in ‘x’ axis and num in ‘y’ axis. Then we compute results using cluster values.

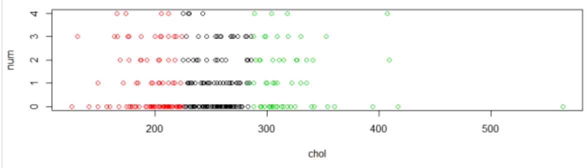


Figure.3.1.K-means clutsring result plot with chol and num

Here in Fig 1.2 we have a plotted a plotting it with similar coordinates against num. Then we get a similar graph. Thus by k-means clustering we can say that num and chol attributes can be used in a better way to predict heart diseases.

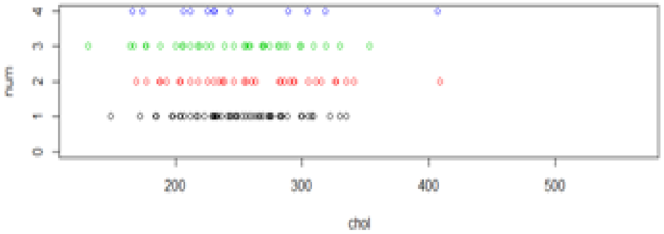
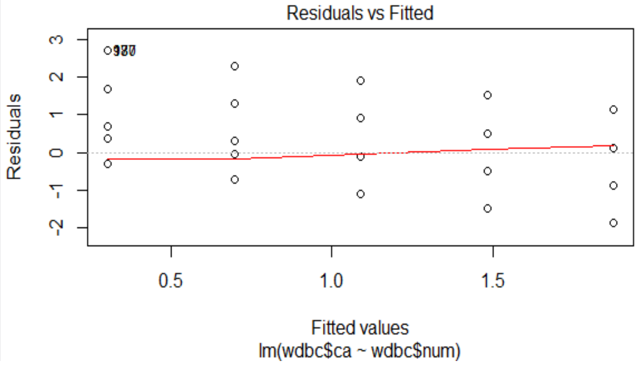
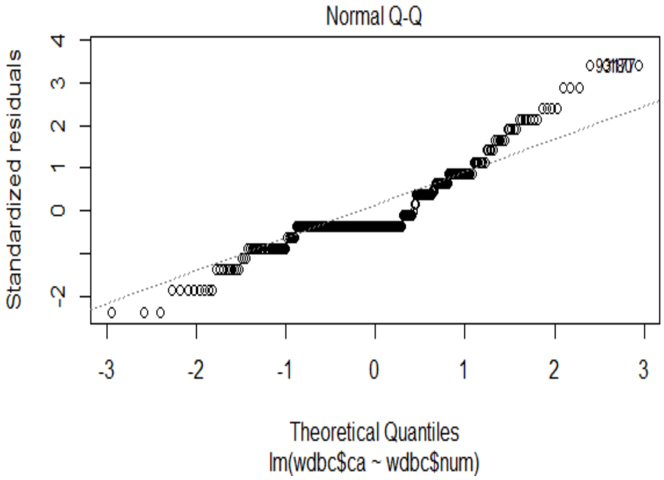


Figure.3.1.K-means clutsring result plot 2

Linear Regression:

To describe the relation between the predictor variables and response variable we use linear regression here. In this technique we used num as dependent variable and any of the other 13 attributes are taken as an independent variable and tested each time. For each and every instance the result is taken as a plot. The result is evaluated using an output "residuals" which gives the difference between predicted and experimental signals. Here the residual errors are randomly and normally distributed and thus this technique is suitable to predict the heart disease. To evaluate the model four graphical approaches are used. The graphs are given below:





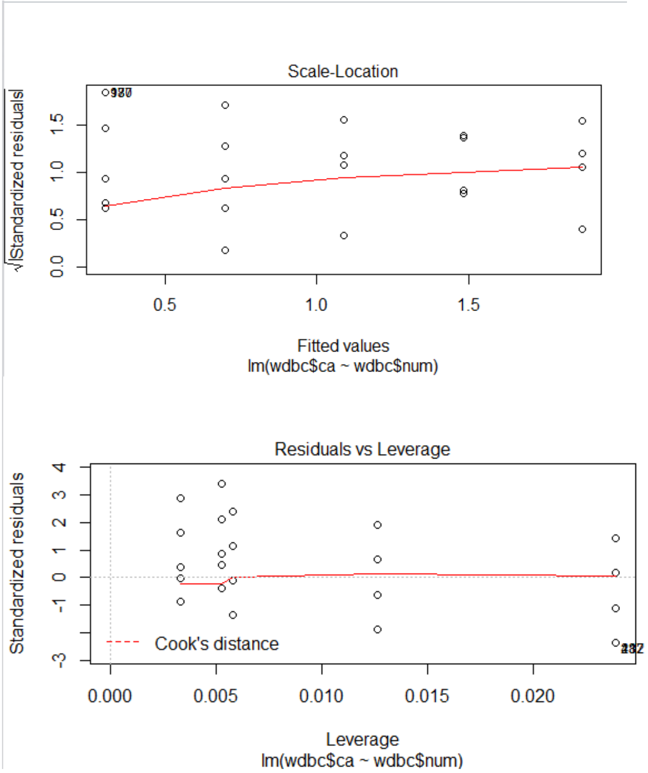


Figure.4.Linearregression residual plots

Output of linear regression:

```
Residuals:
Min   1Q  Median   3Q     Max
-1.8734 -0.3047 -0.3047  0.5188  2.6953

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.30466     0.05771    5.279 2.49e-07 ***
wdbc$num      0.39217     0.03739  10.489  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7982 on 301 degrees of freedom
Multiple R-squared:  0.2677, Adjusted R-squared:  0.2653
F-statistic:   110 on 1 and 301 DF, p-value: < 2.2e-16
```

VIII. CONCLUSION AND FUTURE WORK:

In this paper the heart disease dataset of Cleveland is taken and subjected to various classification and clustering algorithms using Weka and R-programming. The main focus is to target all possible combinations of the attributes against various algorithms. Then of all the techniques it is the technique that works the best to predict the heart disease at an early stage is identified. In weka it is observed that among these algorithms random tree classifier works the best giving good results. From the observation of these results ,we conclude that the heart diseases can be predicted at an early stage using Random Forest K-NN classifiers ,Simple K-means clustering. It is also observed that the attributes like cp(chest pain), exang(exercise induced angina), slope (slope of the peak exercise ST segment), restecg (resting electro cardio graphic results) have greater influence in predicting the heart disease This work can be further expanded in future to predict the heart disease using automation, Artificial Intelligence can also be used to get optimum accuracy in prediction of the heart disease.

IX. REFERENCES:

[1]. Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, *17*(8), 43-48.

[2]. Chaurasia, V., & Pal, S. (2013). Early prediction of heart diseases using data mining techniques. *Caribbean Journal of Science and Technology*, *1*, 208-217.

[3]. Thenmozhi, K., & Deepika, P. (2014). Heart disease prediction using classification with different decision tree techniques. *International Journal of Engineering Research and General Science*, *2*(6), 6-11.

[4]. Abdar, M., Kalhori, S. R. N., Sutikno, T., Subroto, I. M. I., & Arji, G. (2015). Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases. *International Journal of Electrical and Computer Engineering (IJECE)*, *5*(6), 1569-1576.

[5]. Srinivas, K., Rani, B. K., & Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. *International Journal on Computer Science and Engineering (IJCSE)*, *2*(02), 250-255.

[6]. Devi, S. K., Krishnapriya, S., & Kalita, D. (2016). Prediction of Heart Disease using Data Mining Techniques. *Indian Journal of Science and Technology*, *9*(39).

[7]. Bhatla, N., & Jyoti, K. (2012). An analysis of heart disease prediction using different data mining techniques. *International Journal of Engineering*, *1*(8), 1-4.

[8]. Masethe, H. D., & Masethe, M. A. (2014, October). Prediction of heart disease using classification algorithms. In *Proceedings of the world Congress on Engineering and computer Science* (Vol. 2, p. 2224).

[9]. Shouman, M., Turner, T., & Stocker, R. (2011, December). Using decision tree for diagnosing heart disease patients. In *Proceedings of the Ninth Australasian Data Mining Conference-Volume 121* (pp. 23-30). Australian Computer Society, Inc..

[10]. Dangare, C. S., & Apte, S. S. (2012). Improved study of heart disease prediction system using data mining classification techniques. *International Journal of Computer Applications*, *47*(10), 44-48.

[11]. Sowmya, D., Srinidhi, B., & Wahida Banu, S. (2017). HEART DISEASE CLASSIFICATION AND ANALYSIS USING R. *HEART DISEASE*, *3*(4).

[12]. Snehapriya, M., & Umadevi, B. A Novel Prediction Approach for Myocardial Infarction Using Data Mining Techniques.

[13]. Nagalakshmi, D., & Balayesu, N. (2017). PREDICTING HEART DISEASE USING DATAMINING TECHNI QUES. *IJITR*, *5*(4), 7036-7037.

[14]. Patel, A., Gandhi, S., Shetty, S., & Tekwani, B. (2017). Heart Disease Prediction Using Data Mining.

[15]. Sultana, M., Haider, A., & Uddin, M. S. (2016, September) Analysis of data mining techniques for heart disease prediction. In *Electrical Engineering and Information Commun ication Technology (ICEEICT), 2016 3rd International Conference on* (pp. 1-5). IEEE.