Carib.J.Sci.Tech

Early Prediction of Heart Diseases Using Data Mining Techniques

Authors & Affiliation:

Vikas Chaurasia
Research Scholar, Sai Nath
University, Ranchi, Jharkhand,
India.

Saurabh Pal
Head, Dept. of MCA,
VBS Purvanchal University,
Jaunpur, India

Correspondence To:

Vikas Chaurasia

Keywords:

Heart disease, Survivability,
Data Mining, CART, ID3,
Decision table.

ISSN 0799-3757

http://caribjscitech.com/

ABSTRACT

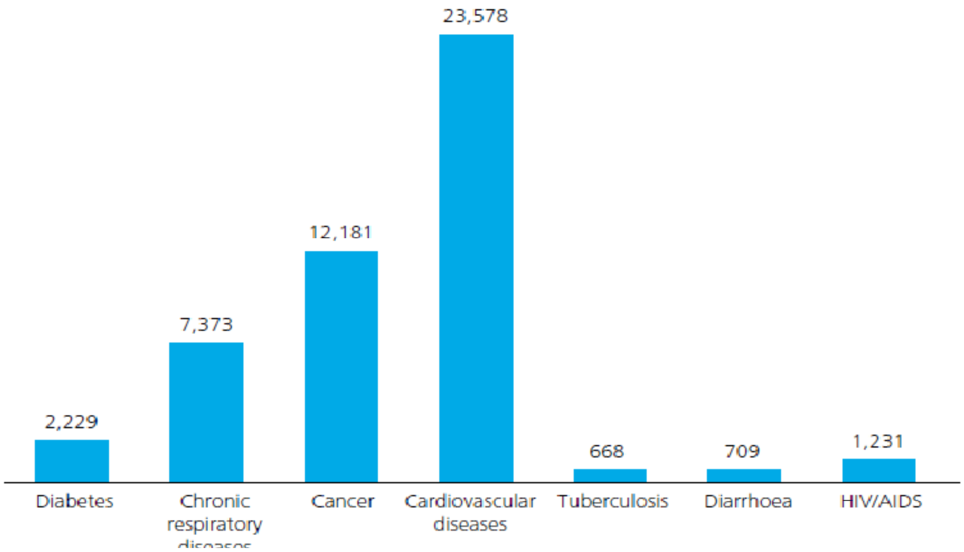
Largest-ever study of deaths shows heart diseases have emerged as the number one killer in world. About 25 per cent of deaths in the age group of 25- 69 years occur because of heart diseases. If all age groups are included, heart diseases account for about 19 per cent of all deaths. It is the leading cause of death among males as well as females. It is also the leading cause of death in all regions though the numbers vary. The proportion of deaths caused by heart disease is the highest in south India (25 per cent) and lowest - 12 per cent - in the central region of India.

The prediction of heart disease survivability has been a challenging research problem for many researchers. Since the early dates of the related research, much advancement has been recorded in several related fields. Therefore, the main objective of this manuscript is to report on a research project where we took advantage of those available technological advancements to develop prediction models for heart disease survivability.

We used three popular data mining algorithms CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and decision table (DT) extracted from a decision tree or rule-based classifier to develop the prediction models using a large dataset. We also used 10-fold cross-validation methods to measure the unbiased estimate.

Introduction

According to a recent study by the Registrar General of India (RGI) and the Indian Council of Medical Research (ICMR), about 25 percent of deaths in the age group of 25- 69 years occur because of heart diseases. In 2008, five out of the top ten causes for mortality worldwide, other than injuries, were non-communicable diseases; this will go up to seven out of ten by the year 2030. By then, about 76% of the deaths in the world will be due to non-communicable diseases (NCDs) [1]. Cardiovascular diseases (CVDs), also on the rise, comprise a major portion of non-communicable diseases. In 2010, of all projected worldwide deaths, 23 million are expected to be because of cardiovascular diseases. In fact, CVDs would be the single largest cause of death in the world accounting for more than a third of all deaths [2].



Source: Global Burden of Diseases 2004. Projected Deaths 2030, Baseline Scenario. World Health Organization, 2008. Number of deaths in '000s

Figure 1: Mortality from major communicable and non-communicable diseases, 2030

Cardiovascular disease includes coronary heart disease (CHD), cerebrovascular disease (stroke), Hypertensive heart disease, congenital heart disease, peripheral artery disease, rheumatic heart disease, inflammatory heart disease. The major causes of cardiovascular disease are tobacco use, physical inactivity, an unhealthy diet and harmful use of alcohol [3]. Several researchers are using statistical and data mining tools to help health care professionals in the diagnosis of heart disease [4].

Complex data mining benefits from the past experience and algorithms defined with existing software and packages, with certain tools gaining a greater affinity or reputation with different techniques [5]. This technique is routinely use in large number of industries like engineering, medicine, crime analysis, expert prediction, Web mining, and mobile computing, besides others utilize Data mining [6]. Medical diagnosis is regarded as an important yet complicated task that needs to be executed accurately and efficiently. The automation of this system would be extremely advantageous. Data mining is an essential step of knowledge discovery. In recent years it has attracted great deal of interest in Information industry [7]. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases [8]. Data mining uses two strategies: supervised and unsupervised learning. In supervised learning, a training set is used to learn model parameters whereas in unsupervised learning no training set is used. Each data mining technique serves a different purpose depending on the modeling objective. The two most common modeling objectives are classification and prediction. Classification models predict categorical labels (discrete, unordered) while prediction models predict continuous-valued functions [9]. Several data mining techniques are used in the diagnosis of heart disease such as Naïve Bayes, Decision Tree, neural network, kernel density, bagging algorithm, and support vector machine showing different levels of accuracies.

This paper presents a new model that enhances the Decision Tree accuracy in identifying heart disease patients. Decision Tree algorithms include CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and C4.5. These algorithms differ in selection of splits, when to stop a node from splitting, and assignment of class to a non-split node (Ho T. J., 2005). The rest of the paper is divided as follows: the background section investigates applying data mining techniques in the diagnosis of heart disease, the methodology section explains the proposed methodology for enhancing the Decision Tree accuracy in diagnosing heart disease, and the results section is followed by a summary section.

Literature Review

Data mining has been played an important role in the intelligent medical systems [11, 12]. The relationships of disorders and the real causes of the disorders and the effects of symptoms that are spontaneously seen in patients can be evaluated by the users via the constructed software easily. Large databases can be applied as the input data to the software by using the extendibility of the software. The effects of relationships that have not been evaluated adequately have been explored and the relationships of hidden knowledge laid among the large medical databases have been searched in this study by means of finding frequent items using candidate generation. The sets of sicknesses simultaneously seen in the medical databases can be reduced by using our non-candidate approach.

Knowledge of the risk factors associated with heart disease helps health care professionals to identify patients at high risk of having heart disease. Statistical analysis and data mining techniques to help healthcare professionals in the diagnosis of heart disease. Statistical analysis has identified the disorders of the heart and blood vessels, and includes coronary heart disease (heart attacks), cerebrovascular disease (stroke), raised blood pressure (hypertension), peripheral artery disease, rheumatic heart disease, congenital heart disease and heart failure. The major causes of cardiovascular disease are tobacco use, physical inactivity, an unhealthy diet and harmful use of alcohol. The three major causes of heart diseases are chest pain, stroke and heart attack [13].

The data mining methods like artificial neural network technique is used in effective heart attack prediction system. First the dataset used for prediction of heart diseases was pre-processed and clustered by means of K-means clustering algorithm [14]. Then neural network is trained with the selected significant patterns. Multi-layer Perceptron Neural Network with Back-propagation is used for training. The results indicate that the algorithm used is capable of predicting the heart diseases more efficiently. The prediction of heart diseases significantly uses 15 attributes, with basic data mining technique like ANN, Clustering and Association Rules, soft computing approaches etc. The outcome shows that Decision Tree performance is more and few times Bayesian classification is having similar accuracy as of decision tree but other predictive methods like K-Nearest Neighbor, Neural Networks, Classification based on clustering will not perform well [15]. By using the Weighted Associative Classifier (WAC), a slight change has been made, instead of considering 5 class labels, only 2 class labels are used. One for “Heart Disease” and another one for “No Heart Disease”. The maximum accuracy (81.51%) has been achieved. When genetic algorithm is applied, the accuracy of the Decision Tree and Bayesian Classification is improved by reducing the actual data size. The dataset of 909 patient records with heart diseases has been collected and 13 attributes has been used for consistency [13]. The patient records have been splitted equally as 455 records for training dataset and 454 records for testing dataset. After applying genetic algorithm the attributes has been reduced to 6 and decision tree performs more efficiently with 99.2% accuracy when compared with other algorithms.

In 2011, Hnin Wint Khaing presented an efficient approach for the prediction of heart attack risk levels from the heart disease database. Firstly, the heart disease database is clustered using the K-means clustering algorithm, which will extract the data relevant to heart attack from the database. This approach allows mastering the number of fragments through its k parameter. Subsequently the frequent patterns are mined from the extracted data, relevant to heart disease, using the MAFIA (Maximal Frequent Item set Algorithm) algorithm. The machine learning algorithm is trained with the selected significant patterns for the effective prediction of heart attack. They have employed the ID3 algorithm as the training algorithm to show level of heart attack with the decision tree. The results showed that the designed prediction system is capable of predicting the heart attack effectively [16].

Chaurasia and Pal conducted study on the prediction of heart attack risk levels from the heart disease database. The prediction of heart diseases significantly uses 11 important attributes, with basic data mining technique like Naïve Bayes, J48 decision tree and Bagging approaches. The outcome shows that bagging techniques performance is more accurate than Bayesian classification and J48. The results shows that the bagging prediction system is capable of predicting the heart attack effectively [17].

Researchers have been applying various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., to help health care professionals with improved accuracy in the diagnosis of heart disease. The heart disease database used from the University of California Irvine. UCI archive is used. This database contains four data sets from the Cleveland Clinic Foundation, Hungarian Institute of Cardiology, V.A. Medical Center and University Hospital of Switzerland. However, here we discuss the Cleveland Heart Disease Dataset (CHDD).

Data Mining Techniques

This paper uses data mining algorithms CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and decision table (DT). These classification algorithms are selected because they are very often used for research purposes and have potential to yield good results. Moreover, they use different approaches for generating the classification models, which increases the chances for finding a prediction model with high classification accuracy.

CART

Decision trees are powerful classification algorithms that are becoming increasingly more popular with the growth of data mining in the field of information systems. Popular decision tree algorithms include Quinlan's ID3, C4.5, C5 [18, 19], and Breiman et al.'s CART [20]. As the name implies, this technique recursively separates observations in branches to construct a tree for the purpose of improving the prediction accuracy. CART builds classification and regression trees for predicting continuous dependent variables (regression) and categorical predictor variables (classification). The classic CART algorithm was popularized by Breiman et al. (Breiman, Friedman, Olshen, & Stone, 1984; see also Ripley, 1996). Although researchers are investigating enhancing CART performance in classification problems, less research is done on enhancing CART performance in disease diagnosis especially in diagnosis of heart disease. In this paper, existing method (CART) is applied to detect heart disease which takes more time and more memory to produce the result. CART(Classification and Regression Tree) uses Gini index to measure the impurity of a partition or set of training tuples. It can handle high dimensional categorical data. Decision Trees can also handle continuous data (as in regression) but they must be converted to categorical data. The CART decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). It can handle high dimensional categorical data. In most cases, the interpretation of results summarized in a tree is very simple. This simplicity is useful not only for purposes of rapid classification of new observations (it is much easier to evaluate just one or two logical conditions, than to compute classification scores for each possible group, or predicted values, based on all predictors and using possibly some complex nonlinear model equations), but can also often yield a much simpler "model" for explaining why observations are classified or predicted in a particular manner. The final results of using tree methods for classification or regression can be summarized in a series of (usually few) logical if-then conditions (tree nodes). Therefore, there is no implicit assumption that the underlying relationships between the predictor variables and the dependent variable are linear, follow some specific non-linear link function, or that they are even monotonic in nature. Thus, tree methods are particularly well suited for data mining tasks, where there is often little a priori knowledge nor any coherent set of theories or predictions regarding which variables are related and how. In those types of data analyses, tree methods can often reveal simple relationships between just a few variables that could have easily gone unnoticed using other analytic techniques.

ID3

Itemized Dichotomized 3 algorithm or better known as ID3 algorithm [18] was first introduced by J.R Quinlan in the late 1970's. The concept of information theory is applied in the field of data mining. As in the algorithms of data mining, the classification is an essential step, using an information theoretic measure in ID3 algorithm, one of the key algorithms of decision tree algorithms, they have discussed the different steps of the development of decision tree so that the best classification criteria can be developed which is helpful in making good decisions. From the data under consideration having a set of values, a property on the basis of calculation is selected as the root of the tree and the process is repeated to develop complete decision tree. ID3, Iterative Dichotomized 3 is a decision tree learning algorithm which is used for the classification of the objects with the iterative inductive approach. In this algorithm the top to down approach is used. The top node is called as the root node and others are the leaf nodes. So it's a traversing from root node to leaf nodes. Each node requires some test on the attributes which decide the level of the leaf nodes. These decision trees are mostly used for the decision making purpose [21, 22]. Data mining techniques basically use the ID3 algorithm as it's the basic algorithm of classification. In the medical field ID3 were mainly used for the data mining.

Decision Table

Decision tables (DTs) provide an alternative way of representing rule-based classification models which is known as tabular representation used to describe and analyze decision situations. Decision tables are easy to interpret and explain to virtually all users, there has been little study of whether such simple models are powerful enough to use for data mining. Much research has concentrated on abstracting accurate models for prediction (or classification) from a given data set. However, a sophisticated technique may remain unused if the model it derives is not comprehensible. For a data mining technique to really be useful, the resulting models should be explainable as well as accurate. Many decisions for example medical treatments cannot be made based on predictions only. Easily interpretable models can give users confidence in the results obtained. The choice of a model affects not only accuracy but also users understanding and confidence in the results.

Heart Disease Data

The data used in this study is the Cleveland Clinic Foundation. Heart disease data set available at http://archive.ics.uci.edu/ml/datasets/Heart+Disease. The data set has 76 raw attributes. However, all of the published experiments only refer to 11 of them. Consequently, to allow comparison with the literature, we restricted testing to these same attributes (see Table I). The data set contains 303 rows.

Table I: SELECTED CLEVELAND CLINIC FOUNDATION

Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	1 = male 0 = female
Cp	Discrete	Chest pain type: 1 = typical angina 2 = atypical angina 3 = non-anginal pa 4 =asymptomatic
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar > 120 mg/dl: 1 = true 0 = false
Restecg	Discrete	Resting electrocardiographic results: 0 = normal 1 = having ST-T wave abnormality 2 =showing probable or define left ventricular hypertrophy by Estes’criteria
Thalach	Continuous	Maximum heart rate achieved
Exang	Discrete	Exercise induced angina: 1 = yes 0 = no
Slope	Discrete	The slope of the peak exercise segment : 1 = up sloping 2 = flat 3= down sloping
Diagnosis	Discrete	Diagnosis classes: 0 = healthy 1= possible heart disease

Data Mining Model

In the describe survey CART, ID3 and decision table have been used to predict attributes such as age, sex, blood pressure and blood sugar for chances of a patient getting heart disease. The data is analyzed and implemented in WEKA ("Waikato Environment for Knowledge Analysis") tool. It is open source software which consists of a collection of machine learning algorithms for data mining tasks. Data mining finds out the valuable information hidden in huge volumes of data. Weka tool is a collection of machine learning algorithms for data mining techniques, written in Java. We have used 10 folds cross validation to minimize any bias in the process and improve the efficiency of the process. The three classifiers like CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and decision table (DT) were implemented in WEKA. The results show clearly that the proposed method performs well compared to other similar methods in the literature, taking into the fact that the attributes taken for analysis are not direct indicators of heart disease.

Results

Here, we analyze heart data set visually using different attributes and figure out the distribution of values. Figure 2 shows the distribution of values of Heart disease patients.

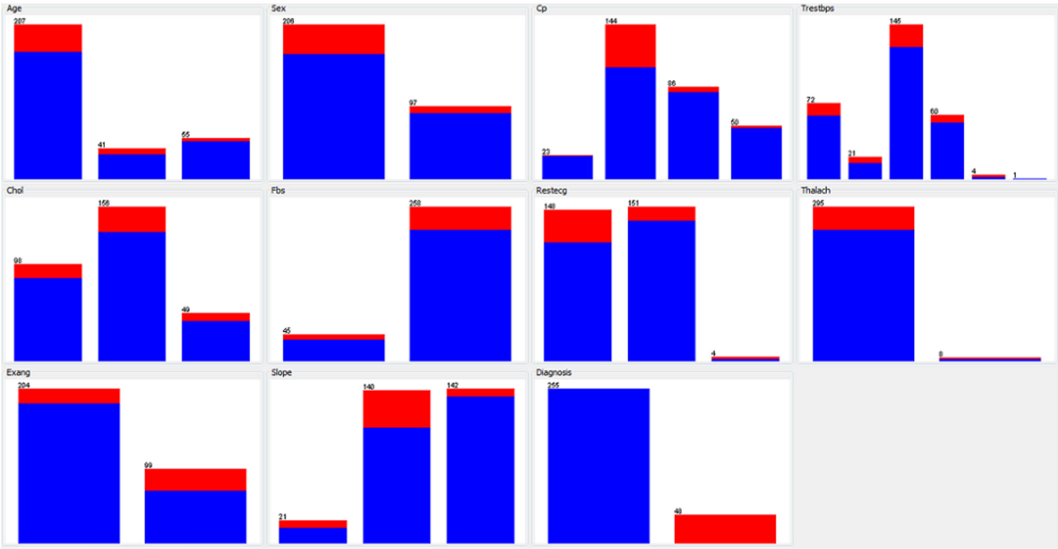


Figure 2: Visualization of the Heart Patients

Table II shows the experimental result. We have carried out some experiments in order to evaluate the performance and usefulness of different classification algorithms for predicting heart patients.

Table II: Performance of the classifiers

Evaluation Criteria	**Classifiers**		
	CART	**ID3**	**Decision Table (DT)**
Timing to build model (in Sec)	0.23	0.02	0.03
Correctly classified instances	253	221	250
Incorrectly classified instances	50	75	53
Accuracy (%)	83.49%	72.93%	82.50%

Here we can show that CART classifier has more accuracy than other classifiers. The percentage of correctly classified instances is often called accuracy or sample accuracy of a model. Kappa statistic, mean absolute error and root mean squared error will be in numeric value only. We also show the relative absolute error and root relative squared error in percentage for references and evaluation. The results of the simulation are shown in Tables III.

Table III: Training and Simulation Error

Evaluation Criteria	Classifiers		
	CART	ID3	Decision Table DT
Kappa statistic(KS)	0.0144	0.0261	-0.0041
Mean absolute error(MAE)	0.2364	0.2669	0.2664
Root mean squared error (RMSE)	0.3448	0.4998	0.3608
Relative absolute error (RAE)	88.07%	103.46%	99.25%
Root relative squared error (RRSE)	94.41%	140.76%	98.78%

Comparison of detailed accuracy by class is shown in table IV.

Table IV: COMPARISON OF ACCURACY MEASURES

Classifier	TP	FP	Precision	Recall	Class
CART	0.988	0.979	0.843	0.988	Healthy
	0.021	0.012	0.25	0.021	Possible Heart Disease
ID3	0.849	0.822	0.852	0.849	Healthy
	0.178	0.151	0.174	0.178	Possible Heart Disease
Decision table (DT)	0.976	0.979	0.841	0.976	Healthy
	0.021	0.024	0.143	0.021	Possible Heart Disease

The performance of the learning techniques is highly dependent on the nature of the training data. Confusion matrices are very useful for evaluating classifiers. The columns represent the predictions, and the rows represent the actual class. To evaluate the robustness of classifier, the usual methodology is to perform cross validation on the classifier.

Table V: Confusion Matrix

Classifier	Healthy	Possible Heart Disease	Class
CART	252	3	Healthy
	47	1	Possible Heart Disease
ID3	213	38	Healthy
	37	8	Possible Heart Disease
Decision table (DT)	249	6	Healthy
	47	1	Possible Heart Disease

Figures 3 and 4 are the graphical representations of the simulation result.

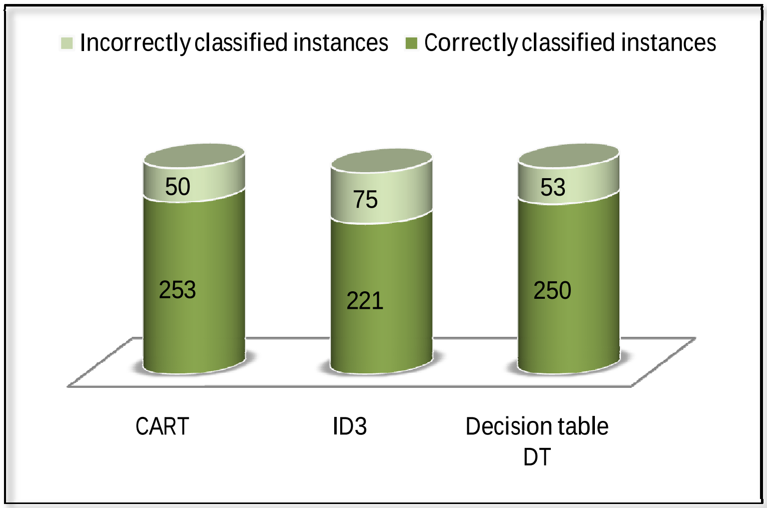


Figure 3: Efficiency of different models

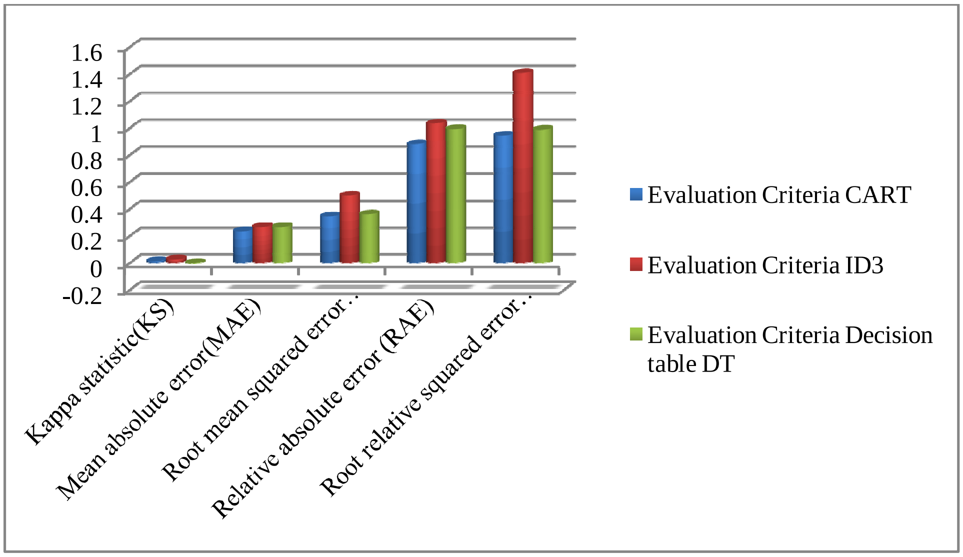


Figure 4: Comparison between Parameters

Based on the above Figures 3, 4 and Table II, we can clearly see that the highest accuracy is 83.49% and the lowest is 72.93%. The other algorithm yields an accuracy of 82.50%. In fact, the highest accuracy belongs to the CART Classifier. An average of 241 instances out of total 303 instances is found to be correctly classified with highest score of 253 instances compared to 221 instances, which is the lowest score. The total time required to build the model is also a crucial parameter in comparing the classification algorithm.

In this simple experiment, from Table II, we can say that a ID3 and DT requires the shortest time which is around 0.02 and 0.03 seconds consecutive with compared to CART which requires the longest model building time which is around 0.23 seconds.

Kappa statistic is used to assess the accuracy of any particular measuring cases, it is usual to distinguish between the reliability of the data collected and their validity. The average Kappa score from the selected algorithm is around -0.0041 - 0.0261. From Figure 4, we can observe the differences of errors resultant from the training of the three selected algorithms. This experiment implies a very commonly used indicator which is mean of absolute errors and root mean squared errors. Alternatively, the relative errors are also used. Since, we have two readings on the errors, taking the average value will be wise.

To better understand the importance of the input variables, it is customary to analyze the impact of input variables during heart disease prediction, in which the impact of certain input variable of the model on the output variable has been analyzed. Tests were conducted using three tests for the assessment of input variables: Chi-square test, Info Gain test and Gain Ratio test. Different algorithms provide very different results, i.e. each of them accounts the relevance of variables in a different way. The average value of all the algorithms is taken as the final result of variables ranking, instead of selecting one algorithm and trusting it. The results obtained with these values are shown in Table VI.

TABLE VI: RESULT OF TESTS AND AVERAGE RANK

Variable	Chi-squared	Info Gain	Gain Ratio	Average Rank
Age	**3.8163**	**0.010515**	**0.008669**	1.278495
Sex	**4.6098**	**0.011886**	**0.01314**	1.544942
Cp	**29.4128**	**0.075006**	**0.043181**	9.843662
Trestbps	**7.2635**	**0.014542**	**0.007901**	2.428648
Chol	**0.2662**	**0.000643**	**0.000445**	0.089096
Fbs	**0.6855**	**0.001548**	**0.002554**	0.229867
Restecg	**12.0985**	**0.027407**	**0.025189**	4.050365
Thalach	**2.8911**	**0.005409**	**0.030729**	0.975746
Exang	**19.9564**	**0.044587**	**0.048913**	6.6833
Slope	**24.9893**	**0.064523**	**0.049863**	8.367895

The aim of this analysis is to determine the importance of each variable individually. Table VI shows that attribute *cp (Chest pain)* impacts output the most, and that it showed the best performances in all of the three tests. Then these attributes follow: *slope (The slope of the peak exercise segment)*, *Exang (Exercise induced angina)*, and *Restecg (Resting electrocardiographic)*. Figure 5 shows the importance of each attributes.

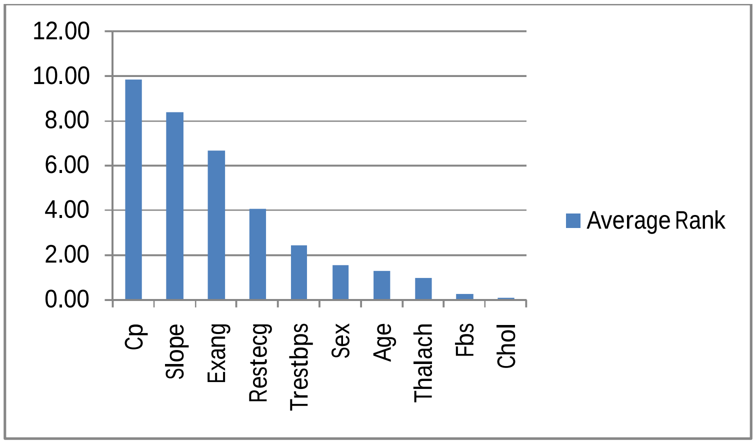


Figure 5: Comparison between importance of attributes

Conclusion

In this paper, different classifiers are studied and the experiments are conducted to find the best classifier for predicting the patient of heart disease.we propose an approach to predict the heart diseases using data mining techniques. Three classifiers such as ID3, CART and DT were used for diagnosis of patients with heart diseases. Observation shows that CART performance is having more accuracy, when compared with other two classification methods.

The best algorithm based on the patient's data is CART Classification with accuracy of 83.49% and the total time taken to build the model is at 0.23 seconds. CART classifier has the lowest average error at 0.3 compared to others. These results suggest that among the machine learning algorithm tested, CART classifier has the potential to significantly improve the conventional classification methods used in the study.
We also shows that the most important attributes for heart diseases are *cp (Chest pain)*, *slope (The slope of the peak exercise segment)*, *Exang (Exercise induced angina)*, and *Restecg (Resting electrocardiographic)*. These attributes were found using three tests for the assessment of input variables: Chi-square test, Info Gain test and Gain Ratio test.

The empirical results show that we can produce short but accurate prediction list for the heart patients by applying the predictive models to the records of incoming new patients. This study will also work to identify those patients who needed special attention.

References

1. Preventing Chronic Disease: A Vital Investment. World Health Organization Global Report, 2005.
2. Global Burden of Disease. 2004 update (2008). World Health Organization.
3. Srinivas, K.," Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques", IEEE Transaction on Computer Science and Education (ICCSE), p(1344 - 1349), 2010.
4. Yanwei Xing, "Combination Data Mining Methods with New Medical Data to Predicting Outcome of Coronary Heart Disease", IEEE Transactions on Convergence Information Technology, pp(868 – 872), 21-23 Nov. 2007
5. IBM, Data mining techniques, http://www.ibm.com/developerworks/opensource/library/ba-data-miningtechniques/ index.html?ca=drs- , downloaded on 04 April 2013.
6. Microsoft Developer Network (MSDN). http://msdn2.microsoft.com/enus/virtuallabs/aa740409.aspx, 2007.
7. Glymour C., D. Madigan, D. Pregidon and P.Smyth, "Statistical inference and data mining", Communication of the ACM, pp: 35-41, 2006.
8. Thuraisingham, B.: "A Primer for Understanding and Applying Data Mining", IT Professional, 28-31, 2000.
9. Han, J., Kamber, M.: "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006.
10. Ho, T. J.: "Data Mining and Data Warehousing", Prentice Hall, 2005.
11. C. Aflori, M. Craus, "Grid implementation of the Apriori algorithm Advances in Engineering Software", Volume 38, Issue 5, May 2007, pp. 295-300.
    A. J.T. Lee, Y.H. Liu, H.Mu Tsai, H.-Hui Lin, H-W. Wu, "Mining frequent patterns in image databases with 9D-SPA representation",Journal of Systems and Software, Volume 82, Issue 4, April 2009, pp.603-618.
12. Srinivas, K., "Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques", IEEE Transaction on Computer Science and Education (ICCSE), p(1344 - 1349), 2010.
13. Shanta kumar, B.Patil,Y.S.Kumaraswamy, "Predictive data mining for medical diagnosis of heart disease prediction" IJCSE Vol .17, 2011
14. M. Anbarasi et. al. "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm", International Journal of Engineering Science and Technology Vol. 2(10), 5370-5376 ,2010.
15. Hnin Wint Khaing, "Data Mining based Fragmentation and Prediction of Medical Data", IEEE, 2011.
16. V. Chauraisa and S. Pal, "Data Mining Approach to Detect Heart Diseases", International Journal of Advanced Computer Science and Information Technology (IJACSIT),Vol. 2, No. 4,2013, pp 56-66.
17. Quinlan J. Induction of decision trees. Mach Learn 1986; 1:81—106.
18. Quinlan J. C4.5: programs for machine learning. San Mateo, CA: Morgan Kaufmann; 1993.
19. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Monterey, CA: Wadsworth & Brooks/ Cole Advanced Books & Software; 1984.
20. Anand Bahety, " Extension and Evaluation of ID3 – Decision Tree Algorithm". University of Maryland, College Park.
21. S. K. Yadav and Pal S., "Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification", World of Computer Science and Information Technology (WCSIT), 2(2), 51-56, 2012.