University Journal of Surgery and Surgical Specialities

ISSN 2455-2860 2018, Vol. 4(4)

# APPLICATION OF DECISION TREE CLASSIFIERS IN DIAGNOSING HEART DISEASE USING DEMOGRAPHIC DATA

***Dr P.Amirtharaj , Dr.Vaithiyanathan V, Rajeswari K**

*Associate Professor, Department of Cardiothoracic Surgery, Madras Medical College, Chennai.

Abstract

The increase in sudden deaths due to heart disease is overwhelming. The decrease in the number of expert doctors has lead to this problem to a large extent. In recent years, machine learning methods are widely used for predictive analysis in Medical diagnosis. This paper proposes a medical decision support system for heart disease risk classification using Decision Tree Classifier. Standard Benchmark database of heart disease from University of California, Irvine (UCI), literature from journals and expert discussions have contributed in designing the attribute set for Ischemic Heart disease for a demographic population in India. The data set is collected from cardio thoracic department, Madras Medical College, Chennai, India from January to April 2011. A total of 712 patients examined are classified into 4 classes using C 4.5 Decision Tree Classifier. Pre-processing steps filtered the data set to 662 patients. The accuracy of correctly classified instances is 82.33%. The same data set with 2 class output gives an accuracy of 92.56%. Whereas the accuracy of UCI Cleveland data set is 78.91%. . The main aim of is to develop more cost-effective and easy-to-use systems, procedures and methods for supporting clinicians.

**Keywords:** Medical decision support system, Ischemic heart disease, pre-processing, Decision Tree Classifier.

## 1. INTRODUCTION

Medical Data mining is a challenging domain as it involves lots of imprecision and uncertainly. The representation of medical knowledge from experts and decision making are real world challenges. Especially in the field of cardiovascular diseases (CVDs), where the expert doctors are not adequate, a clinical decision support system will make a significant contribution. According to WHO [1], heart disease is the leading cause of death and an estimated 17.3 million people died from CVD in 2008. Out of 17.3 million cardio vascular deaths, heart attacks were responsible for 7.3 million deaths and strokes were responsible for 6.2 million deaths. Over 80% of CVD deaths take place in low- and middle-income countries. American Heart Association's Heart Disease and Stroke Statistics of 2012 has estimated that an additional 195,000 silent first myocardial infarctions (heart attacks) occur each year [2, 3].

Heart disease or coronary artery disease (CAD) or coronary heart disease (CHD) or ischemic heart disease (IHD) [10] is a broad term that can refer to any condition that affects the heart [1]. For developing clinical decision support systems, literature presents a number of researches that have made use of artificial intelligence and data mining techniques. Till now, several studies have been reported on heart disease diagnosis. These studies have applied different approaches to the given problem and achieved high classification accuracies, of 77% or higher, using the dataset taken from the UCI machine learning repository. Experimental results [4] showed a correct classification accuracy of approximately 77% with a logistic-regression-derived discriminant function. The John Gennari [5] LAS SIT conceptual clustering system achieved a 78.9% accuracy on the Cleveland database. A Fuzzy Support Vector Clustering to identify heart disease was used in [6]. Resul Das [7] introduced a methodology that uses SAS base software 9.13 for diagnosing heart disease. Zheng Yao [8] applied a new model called R-C4.5 which improved the efficiency of attribution selection and partitioning models. Gang Kou partition [9] applied data separation-based techniques to preserve privacy in the classification of medical data.

CHD have reached epidemic proportions among Indians [14]. India is undergoing a rapid health transition with rising burden of CHD [15] . Further, the long-term case fatality following acute coronary syndrome is considerably higher among Indians as compared to other populations [17]. In addition, a reversal of socio-economic gradients for CHD risk factors has emerged in the Indian population [18, 19]. In this work, we have identified a system for automated medical diagnosis of heart disease risk using decision tree classifier. The success of population based interventions, addressing multiple risk factors for CHDs, through lifestyle linked community programmes was demonstrated initially in North Karelia study2[20].

In developing countries such as India such measures may indeed work due to several reasons. First, the risk factor levels are high among Indians conferring a higher risk. Interventions are likely to have a higher impact on high risk population [14]. CVD are the leading cause of death and disability in both developed and developing countries. A paradigm shift away from the biomedical model is therefore required in the perspective of the existing health care system while responding to the rapidly increasing burden of CVD morbidity and mortality in India [12]. Uneducated and less educated people in rural India have a higher prevalence of coronary heart disease and of the coronary risk factors smoking and hypertension [11].Analysis of data suggest that the risk for CVD and stroke is at epidemic proportions in a cohort of well-educated physicians who are in the highest quintile of income [13]. CVD affects people of all income levels[ll,13].

## 2. MATERIALS AND METHODS

### 2.1 Decision tree Classification Algorithm

Decision tree represents rules which can be easily understood by human and hence used in knowledge mining in databases. It is a classifier with the structure of a tree where each node specifies a test on a single attribute; leaf node indicates the value of the target attribute.

Basic algorithm[26] is a greedy algorithm.

Tree is constructed in a top-down recursive divide-and-conquer manner

At start, all the training examples are at the root

Attributes are categorical (if continuous-valued, they are discretized in advance)

Examples are partitioned recursively based on selected attributes

Test attributes are selected on the basis of a heuristic or statistical measure (e.g., information gain)

The strength of Decision tree is that it generates understandable rules. Implementation of this research included the following steps; patient data collection from Madras Medical College, Preprocessing the data, using Classification and finally this paper compares the results obtained from Cleveland heart disease dataset of UCI and dataset collected from Madras Medical College.

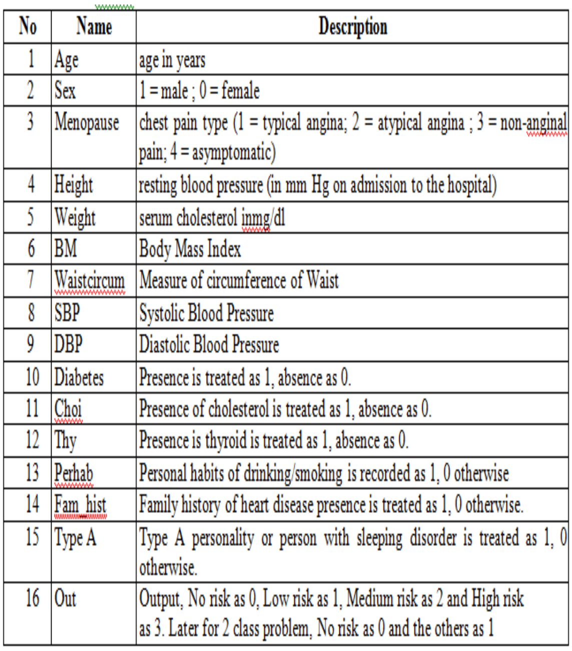
### 2.2. UCI Database description

The heart disease database from the University of California Irvine. UCI archive[21] 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. It provides 920 records in total. Originally, the database had 76 raw attributes. However, all of the published experiments only refer to 13 of these: Age, Sex, P, Trstbps, Choi, Fbs, Restecg, Thalach, Exang, OldPeak, Slope, Ca, Thai and Num.

### 2.3 Madras Medical College database description

The following features are proposed to be collected and analyzed for Indian Heart risk score prediction based on extensive study and expert opinion from doctors with respect to Indian body conditions, life style and eating habits. After discussion with cardiologists a questionnaire was prepared. Diagnosis is done through data collection for each individual patient as given in Table 1.

**Table. 1:** Attributes identified for heart disease identification

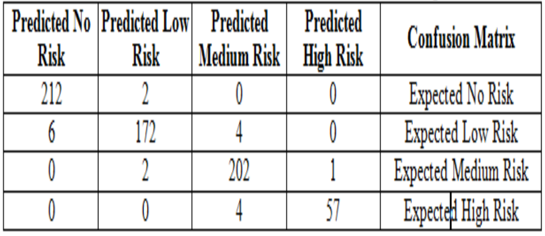


| No | Name | Description |
|---|---|---|
| 1 | Age | age in years |
| 2 | Sex | 1 = male ; 0 = female |
| 3 | Menopause | chest pain type (1 = typical angina; 2 = atypical angina ; 3 = non-anginal pain; 4 = asymptomatic) |
| 4 | Height | resting blood pressure (in mm Hg on admission to the hospital) |
| 5 | Weight | serum cholesterol inmg/dl |
| 6 | BM | Body Mass Index |
| 7 | Waistcircum | Measure of circumference of Waist |
| 8 | SBP | Systolic Blood Pressure |
| 9 | DBP | Diastolic Blood Pressure |
| 10 | Diabetes | Presence is treated as 1, absence as 0. |
| 11 | Choi | Presence of cholesterol is treated as 1, absence as 0. |
| 12 | Thy | Presence is thyroid is treated as 1, absence as 0. |
| 13 | Perhab | Personal habits of drinking/smoking is recorded as 1, 0 otherwise |
| 14 | Fam_hist | Family history of heart disease presence is treated as 1, 0 otherwise. |
| 15 | Type A | Type A personality or person with sleeping disorder is treated as 1, 0 otherwise. |
| 16 | Out | Output, No risk as 0, Low risk as 1, Medium risk as 2 and High risk as 3. Later for 2 class problem, No risk as 0 and the others as 1 |

## 3. RESULTS AND DISCUSSION:

For experimentation, the heart disease datasets collected from Madras Medical college are divided into two sets such as: (1) training dataset and (2) testing dataset using 10-fold cross validation.. The problem of missing values of a particular attribute is not there as data is collected from primary source. Weka; a open source software is used with J48 classifier. The 712 data set when used as training data with 10-fold cross validation, results in the form of confusion matrix is shown in Table 2.

**Table 2:** Confusion Matrix with 10-fold cross validation on training data set



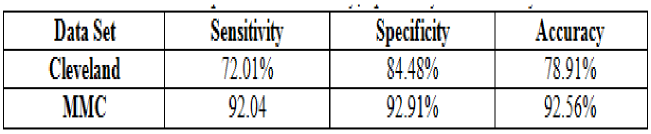
| Predicted No Risk | Predicted Low Risk | Predicted Medium Risk | Predicted High Risk | Confusion Matrix |
|---|---|---|---|---|
| 212 | 2 | 0 | 0 | Expected No Risk |
| 6 | 172 | 4 | 0 | Expected Low Risk |
| 0 | 2 | 202 | 1 | Expected Medium Risk |
| 0 | 0 | 4 | 57 | Expected High Risk |

Time taken during training to build mode is 0.08 seconds. Time taken to test model on training data is 0.02 seconds. In the testing phase, the testing dataset is given to the proposed system to find the risk prediction of heart patients and obtained results are evaluated with the evaluation metrics namely, sensitivity, specificity and accuracy[22]

Sensitivity= TP/(TP+FN); Specificity = TO/ (TN+FP); Accuracy= (TN+TP)/ (TN+TP+FN+FP)

In the UCI data set, value 0 specifies the no presence of heart disease (less than 50% diameter narrowing) and values 1—4 specifies the presence of heart disease (greater the 50% diameter narrowing). According to this, we have transformed it into two class data, where class 0 specifies the no presence of heart disease and class 1 specifies the presence of heart disease. The training dataset is used to generate the rules and the testing dataset is used to analyze the performance of the proposed system.

With the training set, with 10-fold cross validation [23,24,25], the comparison of sensitivity, specificity and accuracy is shown below in Table 3 and Figure 1.



| Data Set | Sensitivity | Specificity | Accuracy |
|---|---|---|---|
| Cleveland | 72.01% | 84.48% | 78.91% |
| MMC | 92.04 | 92.91% | 92.56% |

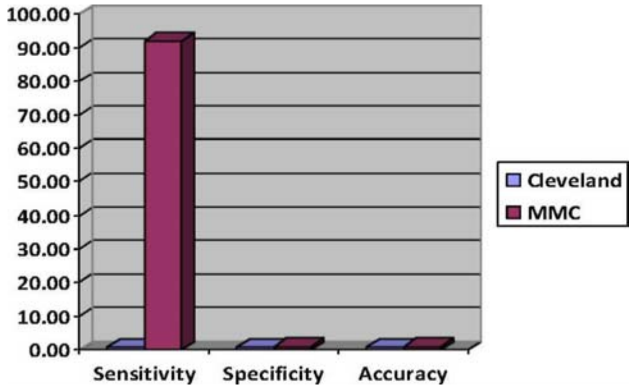


**Figure 1:** Comparison of UCI data and MMC

During testing, 50 patient's data was given. Average prediction was 0.96. Accuracy percentage was 82%.Time taken to build model: 0.08 seconds. Accuracy is number of correctly classified instances, 93.56%.

**4. Conclusion**

Cardiovascular Diseases include Heart Disease and Stroke disease. In this paper Heart Disease is taken for classification problem with C4.5 classifier of Weka. Standard benchmark Cleveland dataset from uci database is compared with the Indian dataset. Accuracy, Sensitivity and Specificity measures are far better for the dataset with modified attributes. These attributes will be very useful for densely populated countries like India where shortage of Expert Doctors is a major problem. The time of Experts can be saved to a larger extent if intelligent decision support system classifies the risk level in preliminary stage. The system may give varied results for different population and hence it has to be verified in future

**References**

http ://www. who. int/cardiovascular_diseases/en/

http://media.trb.com/media/acrobat/2012-01/285933360-30132728.pdf

http ://circ. ahaj ournals. org/content/125/ l/e2 .full

R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.J. Schmid, S. Sandhu, K.H.Guppy, S. Lee, and V. Froelicher. "International application of a new probability algorithm for the diagnosis of coronary artery disease". American Journal of Cardiology, Vol. 64, 1989, pp. 304-310

G. John. "Models if incremental concept formation". Journal of Atificial Intelligence, 1989, pp. 11-61.

L. Gamboa, M. G. Mendoza, J. M. Vargas, N. H. Gress, and R. E. Orozco, "Hybrid Fuzzy-SV Clustering for Heart Disease identification", in Proceedings of CIMCA- IAWTIC'06. 2006.

D. Resul, T. Ibrahim, S. Abdulkadir,. "Effective diagnosis of heart disease through neural Elsevier, 2008.

Z. Yao, P. Liu, L. Lei, and J. Yin, "R-C4.5 Decision tree model and its applications to health care dataset", in networks ensembles". Proceedings of the 2005 International Conference on Services Systems and Services Management. 2005. p. 1099- 1103.

K. Gang, P. Yi, S. Yong, C. Zhengxin,. "Privacy-preserving data mining of medical data using data separation- based techniques". Data science journal, Vol. 6, 2007.

http://heart-disease.emedtv.com/coronary-artery-disease/coronary-artery-disease.html

R. Gupta, V. P. Gupta, and N. S. Ahluwalia, 'Educational status, coronary heart disease, and coronary risk factor prevalence in a rural population of India', BMJ. 1994 November 19; 309(6965): 1332-1336

Panniyammakal Jeemon & K.S. Reddy, 'Social determinants of cardiovascular disease outcomes in Indians', Indian J Med Res 132, November 2010, pp 617-622.

A Mathavan, MD, A Chockalingam, PhD, S Chockalingam, BSc, B Bilchik, MD, and V Saini, MD, Madurai Area Physicians Cardiovascular Health Evaluation Survey (MAPCHES) - an alarming status', The Canadian Journal of Cardiology 2009 May; 25(5): 303-308.

Vamadevan S. Ajay & Dorairaj Prabhakaran, 'Coronary heart disease in Indians: Implications of the INTERHEART study', Indian J Med Res 132, November 2010, pp 561-566.

Reddy KS, Shah B, Varghese C, Ramadoss A. Responding to the threat of chronic diseases in India. Lancet 2005; 366 : 1744-9.

Gupta R, Joshi P, Mohan V, Reddy KS, Yusuf S. Epidemiology and causation of coronary heart disease and stroke in India. Heart 2008; 94 : 16-26

Prabhakaran D, Yusuf S, Mehta S, Pogue J, Avezum A, Budaj A, et al. Two-year outcomes in patients admitted with non- ST elevation acute coronary syndrome: results of the OASIS registry 1 and 2. Indian Heart J 2005; 57 : 217-25

Reddy KS, Prabhakaran D, Jeemon P, Thankappan KR, Joshi P, Chaturvedi V, et al. Educational status and cardiovascular risk profile in Indians. Proc Natl Acad Sci USA 2007; 104 : 16263-8.

Ajay VS, Prabhakaran D, Jeemon P, Thankappan KR, Mohan V, Ramakrishnan L, et al. Prevalence and determinants of diabetes mellitus in the Indian Industrial population. Diabetic Med 2008; 25 : 1187-94.

Puska P, Tuomilehto J, Aulikki N, Enkki V. The North Karelia Project. 20 years results and experiences. Helsinki: National Public Health Institute; 1995.

Newman, D.J., Hettich, S., Blake, C.L., Merz, C.J., 1998. UCI repository of machine learning databases. Department of Information and Computer Science, University California Irvine

Zhu, W., Zeng, N., Wang, N., 2010. Sensitivity, specificity, accuracy, associated confidence interval and roc analysis with practical SAS implementations. In: NESUG Proceedings: Health Care and Life Sciences, Baltimore, Maryland

http://www.public.asu.edu/~ltang9/papers/ency-cross-validation.pdf

McLachlan, Geoffrey J.; Do, Kim-Anh; Ambroise, Christophe (2004). Analyzing microarray gene expression data. Wiley.

Geisser, Seymour (1993). Predictive Inference. New York, NY: Chapman and Hall. ISBN 0-412-03471-9

Jiawei Han, Micheline Kamber, 'Datamining: Concepts and Techniques', 2nd Edition, Morgan Kaufmann Publishers, March 2006. ISBN 1-55860-901-6