



## Prediction of Heart Disease Using Decision Tree Approach

R. Vijaya Kumar Reddy<sup>1</sup>, K. Prudvi Raju<sup>2</sup>, M. Jogendra Kumar<sup>3</sup>, CH. Sujatha<sup>4</sup>, P. Ravi Prakash<sup>5</sup>

<sup>1,2,5</sup> Asst. Professor, Department of IT, PVPSIT, Kanuru, Vijayawada, Andhra Pradesh, India

<sup>3,4</sup> Asst. Professor, Department of ECM, PVPSIT, Kanuru, Vijayawada, Andhra Pradesh, India

---

**Abstract:** *In today's modern world cardiovascular disease is the most lethal one. This disease attacks a person so instantly that it hardly gets any time to get treated with. So diagnosing patients correctly on timely basis is the most challenging task for the medical fraternity. Poor clinical decisions may end to patient death and which cannot be afforded by the hospital as it loses its reputation. In this paper using a data mining technique Decision Tree is used an attempt is made to assist in the diagnosis of the disease, Keeping in view the goal of this study to predict heart disease using classification techniques, I have used a supervised machine learning algorithms i.e., Decision Tree. It has been shown that, by using a decision tree, it is possible to predict heart disease vulnerability in diabetic patients with reasonable accuracy. Classifiers of this kind can help in early detection of the vulnerability of a diabetic patient to heart disease.*

**Keywords:** *Data mining; Decision Tree algorithm; Classification; Predication;*

---

### I. INTRODUCTION

Medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis. However, the available raw medical data are widely distributed, heterogeneous in nature, and voluminous. These data need to be collected in an organized form. This collected data can be then integrated to form a hospital information system. Data mining technology provides a user oriented approach to novel and hidden patterns in the data.

Today diagnosing patients correctly and administering effective treatments have become quite a challenge. Poor clinical decisions may end to patient death and which cannot be afforded by the hospital as it loses its reputation. The cost to treat a patient with a heart problem is quite high and not affordable by every patient. To achieve a correct and cost effective treatment computer-based information and/or decision support Systems can be developed to do the task [1]. Most hospitals today use some sort of hospital information systems to manage their healthcare or patient data. These systems typically generate huge amounts of data which take the form of numbers, text, charts and images. Unfortunately, these data are rarely used to support clinical decision making. There is a wealth of hidden information in these data that is largely untapped. This raises an important question: "How can we turn data into useful information that can enable healthcare practitioners to make intelligent clinical decisions?" So there is need of developing a master's project which will help practitioners predict the heart disease before it occurs. The diagnosis of diseases is a vital and intricate job in medicine [9]. The recognition of heart disease from diverse features or signs is a multi-layered problem that is not free from false assumptions and is frequently accompanied by impulsive effects. Thus the attempt to exploit knowledge and experience of several specialists and clinical screening data of patients composed in databases to assist the diagnosis procedure is regarded as a valuable option.

The World Health Organization has estimated that 12 million deaths occurs worldwide, every year due to the Heart diseases. It is also the chief reason of deaths in numerous developing countries. On the whole, it is regarded as the primary reason behind deaths in adults. The term Heart disease encompasses the diverse diseases that affect the heart. Heart disease was the major cause of casualties in the different countries including India. Coronary heart disease, Cardiomyopathy and Cardiovascular disease are some categories of heart diseases. The term "cardiovascular disease" includes a wide range of conditions that affect the heart and the blood vessels and the manner in which blood is pumped and circulated through the body. Cardiovascular disease (CVD) results in several illness, disability, and death. The diagnosis of diseases is a vital and intricate job in medicine.

### II. LITERATURE SURVEY

Numerous studies have been done that have focus on diagnosis of heart disease. They have applied different data mining techniques for diagnosis & achieved different probabilities for different methods.

N. Deepika et al. proposed Association Rule for classification of Heart-attack patients [1]. The extraction of significant patterns from the heart disease data warehouse was presented. The heart disease data warehouse contains the screening clinical data of heart patients. Initially, the data warehouse preprocessed to make the mining process more efficient. The first stage of Association Rule used preprocessing in order to handle missing values. Later applied equal interval binning with approximate values based on medical expert advice on Pima Indian heart attack data. The

significant items were calculated for all frequent patterns with the aid of the proposed approach. The frequent patterns with confidence greater than a predefined threshold were chosen and it was used in the design and development of the heart attack prediction system.

K. Srinivas et al. presented Application of Data Mining Technique in Healthcare and Prediction of Heart Attacks [2]. The potential use of classification based data mining techniques such as Rule based, Decision tree, Naïve Bayes and Artificial Neural Network to the massive Volume of healthcare data.

Sudha et al. [3] to propose the classification algorithm like Naïve Bayes, Decision tree and Neural Network for predicting the stroke diseases. The classification algorithm like decision trees, Bayesian classifier and back propagation neural network were adopted in this study.

M A. Jabbar et al. proposed Association Rule mining based on the sequence number and clustering for heart attack prediction [4]. The entire database is divided into partitions of equal size. The dataset with 14 attributes was used in that work and also each cluster is considered one at a time for calculating frequent item sets. This approach reduces main memory requirement. To predict the heart attack in an efficient way the patterns are extracted from the database with significant weight calculation. The frequent patterns having a value greater than a predefined threshold were chosen for the valuable prediction of heart attack. Three mining goals were defined based on data exploration and all those models could answer complex queries in predicting heart attack.[7].

### III. ALGORITHM

**Decision Trees** The decision tree approach is more powerful for classification problems. There are two steps in this techniques building a tree & applying the tree to the dataset. There are many popular decision tree algorithms CART, ID3, C4.5, CHAID, and J48. From these J48 algorithm is used for this system. J48 algorithm uses pruning method to build a tree. Pruning is a technique that reduces size of tree by removing over fitting data, which leads to poor accuracy in predications. The J48 algorithm recursively classifies data until it has been categorized as perfectly as possible. This technique gives maximum accuracy on training data. The overall concept is to build a tree that provides balance of flexibility & accuracy.

#### A. preprocessing

The actions comprised in the preprocessing of a data set are the removal of duplicate records, normalizing the values used to represent information in the database, accounting for missing data points and removing unneeded data fields.

Moreover it might be essential to combine the data so as to reduce the number of data sets besides minimizing the memory and processing resources required by the data mining algorithm [5]. In the real world, data is not always complete and in the case of the medical data, it is always true. To remove the number of inconsistencies which are associated with data we use Data preprocessing. The Pre-process panel has facilities for importing data from a database, a comma-separated values (CSV) file, etc., and for pre-processing this data using a so-called filtering algorithm. These filters can be used to transform the data (e.g., turning numeric attributes into discrete ones) and make it possible to delete instances and attributes according to specific criteria.

#### B. Classification

The records with irrelevant data were removed from data warehouse before mining process occurs. Data mining classification technology consists of classification model and evaluation model. The classification model makes use of training data set in order to build classification predictive model. The testing data set was used for testing the classification efficiency. Then the classification algorithm like decision tree, naive Bayes and neural network was used for stroke disease prediction[3]. The performance evaluation was carried out based on Decision Tree algorithms and accuracy was measured. The Classify panel enables applying classification and regression algorithms (indiscriminately called classifiers in Weka) to the resulting dataset, to estimate the accuracy of the resulting predictive model, and to visualize erroneous predictions, receiver operating characteristic (ROC) curves, etc., or the model itself (if the model is amenable to visualization like, e.g., a decision tree).

#### C. Decision tree

Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. It is one of the predictive modelling approaches used in statistics, data mining and machine learning. Tree models where the target variable can take a finite set of values are called classification trees. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data but not decisions; rather the resulting classification tree can be an input for decision making.

#### D. Clustering

The Cluster panel gives access to the clustering techniques in Weka, e.g., the simple k-means algorithm[4]. k-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem.

Let  $X = \{x_1, x_2, x_3, \dots, x_n\}$  be the set of data points and  $V = \{v_1, v_2, \dots, v_c\}$  be the set of centers. Randomly select 'c' cluster centers. Calculate the distance between each data point and cluster centers. Assign the data point to the

cluster center whose distance from the cluster center is minimum of all the cluster centers. Recalculate the new cluster center using:

$$v_i = (1 / c_i) \sum_{j=1}^{c_i} x_j \quad (1)$$

where, ' $c_i$ ' represents the number of data points in  $i^{th}$  cluster. Recalculate the distance between each data point and new obtained cluster centers. If no data point was reassigned then stop, otherwise repeat from assign the data point.

## V. PROPOSED SYSTEM

Our project has been mainly developed with an aim to efficiently diagnose the presence of heart disease in an individual. For this purpose we are going to use JAVA as our front end where in we could create a user interface to accept user details and back end would be MYSQL. The front end would basically work as.

**REGISTER:** Firstly, if the patient is not registered or is arriving for the first time to the –doctor he should register himself so that his information can be stored in the database which would be useful in the future for diagnosis. So Initially the patient needs to register himself for the system. But, if the patient is an old user then he might go for the next step as below.

**LOGIN:** In this step, the patient would login through his user id and access his own profile where in JAVA would be useful for giving access to the patient's profile.

**USER INPUT:** After accessing the profile the doctor would enter the details of the patient as mentioned by him. The doctor would mainly undertake tests considering the of attributes in mind such as Blood Pressure – where in the values observed by the doctor would be entered in the field corresponding to Blood Pressure. Similarly all other values corresponding to the associated attributes such as LDL – Low Density Lipoprotein (commonly known as Bad Cholesterol), HDL – High Density Lipoprotein (commonly known as Good Cholesterol) and Triglycerides observed by the doctor would be entered by him respectively. Thus this would complete all the information required from the patient.

**FINAL REPORT:** After getting the information from the patient, Data mining would be utilized where in the current details of the patient would be compared by his previous details and Decision Tree algorithm would be used to identify if the patient has some symptoms of Heart Diseases or not. Thus, in order to access the patients history MYSQL would also be used as the Back end for our System.

## VI. CONCLUSION

The decision-tree algorithm is one of the most effective and efficient classification methods available. It has been shown that, by using a decision tree, it is possible to predict heart disease vulnerability in diabetic patients with reasonable accuracy. Classifiers of this kind can help in early detection of the vulnerability of a diabetic patient to heart disease. Preprocessing of a data set for the removal of duplicate records, normalizing the values used to represent information in the database. Clustering technique, simple k-means algorithm is used. Thus, the patients can be forewarned to change their lifestyles. This will result in preventing diabetic patients from being affected by heart diseases, thereby resulting in low mortality rates as well as reduced cost on health care for the state. This can be extended in future to predict other types of ailments which arise from diabetes, such as visual impairment. The proposed work can be further enhanced and expanded, to use stacking techniques to increase the accuracy of decision trees and reduce the number of leaf nodes.

## REFERENCES

- [1] N. Deepika and K. Chandra shekar, "Association rule for classification of Heart Attack Patients", *International Journal of Advanced Engineering Science and Technologies*, Vol. 11, No. 2, pp. 253 – 257, 2011.
- [2] K. Srinivas, B. Kavitha Rani and Dr. A. Govrdhan, "Application of Data Mining Techniques in Healthcare and Prediction of Heart Attacks", *International Journal on Computer Science and Engineering*, Vol. 02, No. 02, pp. 250 - 255, 2011.
- [3] A. Sudha, P. Gayathiri and N. Jaisankar, "Effective Analysis and Predictive Model of Stroke Disease using Classification Methods", *International Journal of Computer Applications*, Vol. 43, No. 14, pp. 0975 – 8887, 2012.
- [4] M A. Jabbar, Priti Chandra and B. L. Deekshatulu, "Cluster based association rule mining for heart attack prediction", *Journal of Theoretical and Applied Information Technology*, Vol. 32, No.2, pp. 197 - 201, 2011.
- [5] D. Shanthi, G. Sahoo and N. Saravanan "Input Feature Selection using Hybrid Neuro-Genetic Approach in the Diagnosis of Stroke Disease", *International Journal of Computer Science and Network Security*, Vol. 8, No.12, pp. 99 - 106, 2008.
- [7] K. Srinivas, G. Raghavendra Rao and A. Govardhan, "Survey on prediction of heart morbidity using data mining techniques", *International Journal of Data Mining & Knowledge Management Process*, Vol. 1, No. 3, pp. 14 -34, 2011.
- [8] Chaitrali S. Dangare, Sulabha S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques"; *International Journal of Computer Applications* (0975 – 888) Volume 47– No.10, June 2012.
- [9] T.Georgeena.S. Thomas, Siddhesh.S. Budhkar, Siddhesh.K. Cheulkar, Akshay.B.Choudhary, Rohan Singh"Heart Disease Diagnosis System Using Apriori Algorithm"; *International Journal of Advanced Research in Computer Science and Software Engineering* Volume 5, Issue 2, February 2015.