A Technical Seminar Report

on

Early Prediction of Heart disease using Decision Tree Algorithm

Submitted to CVR College of Engineering

By

CH.HARSHITH

17B81A05R2

As Part of Academic Requirement for B.Tech Degree

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CVR COLLEGE OF ENGINEERING

(***An UGC Autonomous institution, Accredited by NAAC with 'A' Grade***)

Academic Year 2019- 2020

CVR COLLEGE OF ENGINEERING

(*An UGC Autonomous institution, Accredited by NAAC with ‘A’ Grade*)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the technical seminar report titled “*Early Prediction of heart disease using Decision tree algorithm*” is submitted by *Ch.Harshith,* bearing H.T.No:17B81A05R2,as part of the academic requirement of the Graduate Engineering Program in Computer Science and Engineering.

Technical Seminar Coordinator

Head of the Department
Dr . A . Vani Vathsala

ACKNOWLEDGEMENT

I sincerely thank **Dr. Nayanathara K Sattiraju**, Principal, CVR College of Engineering, for her cooperation and encouragement throughout the technical seminar.

I earnestly thank **Dr. A. Vani Vathsala**, Head of Department, Department of Computer Science and Engineering, CVR College of Engineering, for giving timely cooperation and taking necessary action throughout the course of my technical seminar.

I express my sincere thanks and gratitude to my coordinator and internal guide **Dr . R . K . Selva Kumar**, Department of Computer Science and Engineering, CVR College of Engineering, for her valuable help and encouragement throughout the technical seminar

Finally, we thank all those whose guidance helped us in this regard. I place in records my sincere appreciation and indebtedness to my parents and all the lecturers for their understanding and cooperation, without whose encouragement and blessing it would not have been possible to complete this work.

With regards,
Ch.Harshith
17B81A05R2

TABLE OF CONTENTS

S.No **Page No**

1.ABSTRACT

For processing large amounts of data numerous techniques are used. Data Mining is one of the techniques that are used most often. To process these data, Data mining combines traditional data analysis with sophisticated algorithms. Medical data mining is an important area of Data Mining and considered as one of the important research fields due to its application in the healthcare domain. Classification and prediction of medical datasets poses challenges in Medical Data Mining. Heart disease accounts to be the leading cause of death worldwide. It is difficult for medical practitioners to predict the heart attack as it is a complex task that requires experience and knowledge. The health sector today contains hidden information that can be important in making decisions. Data mining algorithms such as decision tree are applied in this research for predicting heart attacks. The research result shows prediction accuracy of 99%. Data mining enables the health sector to predict patterns in the dataset.

Key Words : Data mining, Decision Tree

2.INTRODUCTION

Data Mining is a non-trivial extraction of implicit, previously unknown and potential useful information about data” [1]. In short, it is a process of analyzing data from different perspective and gathering the knowledge from it. The discovered knowledge can be used for different applications for example the healthcare industry. Nowadays healthcare industry generates a large amount of data about patients, disease diagnosis etc. Data mining provides a set of techniques to discover hidden patterns from data. A major challenge facing Healthcare industry is quality of service. Quality of service implies diagnosing disease correctly &

provides effective treatments to patients. Poor diagnosis can lead to disastrous consequences That is unacceptable.Therefore, an automatic medical diagnosis system is designed that takes advantage of a collected database and decision support system. This system can help in diagnosing disease with less medical tests & effective treatments.

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.

3.OBJECTIVE OF SEMINAR

Appropriate computer-based information and/or decision support systems can aid in achieving clinical tests at a reduced cost. Efficient and accurate implementation of automated system needs a comparative study of various techniques available. This paper aims to analyze the different predictive/ descriptive data mining techniques proposed in recent years for the diagnosis of heart disease.Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. The automation of the same would be highly beneficial. Clinical decisions are often made based on doctor’s intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Data mining has the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions. Decision Tree is a popular classifier which is simple and easy to implement. It requires no domain knowledge or parameter setting and can handle high dimensional data. The results obtained from Decision Trees are easier to read and interpret. The drill through feature to access detailed patients‟ profiles is only available in Decision Trees.

4.ORGANIZATION OF THE THESIS

This chapter is organized as follows: first, we outline the basics of patient physiology and fetus response to different stages of oxygen deficiency - hypo anemia, hypoxia, and asphyxia. Next, we describe an interaction between mother and fetus during gestation with emphasis on the antepartum and intrapartum period. Finally, we introduce methods for the patient hypoxia diagnostics with focus on electronic patient monitoring that involves observation of CTG or FECG changes. We stress the significance of signal interpretation and describe advantages and disadvantages of respective methods.

5.ALGORITHM

The decision tree approach is more powerful for classification problems. There are two steps in this technique: 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 algorithms are used for this system. The J48 algorithm uses a pruning method to build a tree. Pruning Is a technique that reduces the size of a tree by removing overfitting data, which leads to poor accuracy in predictions. 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.Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy). Leaf node (e.g., Play) represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data. The constructing decision tree techniques are generally computationally inexpensive, making it possible to quickly construct models even when the training set size is very large. Furthermore, once a decision tree has been built, classifying a test record is extremely fast.

6.CLASSIFICATION USING DECISION TREE ALGORITHM

6.1 ENTROPY :

A decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values (homogenous). The ID3 algorithm uses entropy to calculate the homogeneity of a sample. If the sample is completely homogeneous.The entropy is zero and if the sample is equally divided it has entropy of one. To build a decision tree, we need to calculate two types of entropy using frequency tables as follows:

a) Entropy using the frequency table of one attributes:

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

b) Entropy using the frequency table of two attributes:

$$E(T,X) = \sum_{c \in X} P(c)E(c)$$

6.2 INFORMATION GAIN :

The information gain is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding the attribute that returns the highest information gain (i.e., the most homogeneous branches).

Step 1: Calculate entropy of the target.

Step 2: The dataset is then split on the different attributes. The entropy for each branch is calculated. Then it is added proportionally, to get total entropy for the split. The resulting entropy is subtracted from the entropy before the split. The result is the Information Gain, or decrease in entropy.

$$Gain(T,X) = Entropy(T) - Entropy(T,X)$$

Step 3: Choose the attribute with the largest information gain as the decision node.

Step 4(a): A branch with entropy of 0 is a leaf node.

Step 4(b): A branch with entropy more than 0 needs further splitting.

Step 5: The ID3 algorithm is run recursively on the non-leaf branches, until all data is classified.

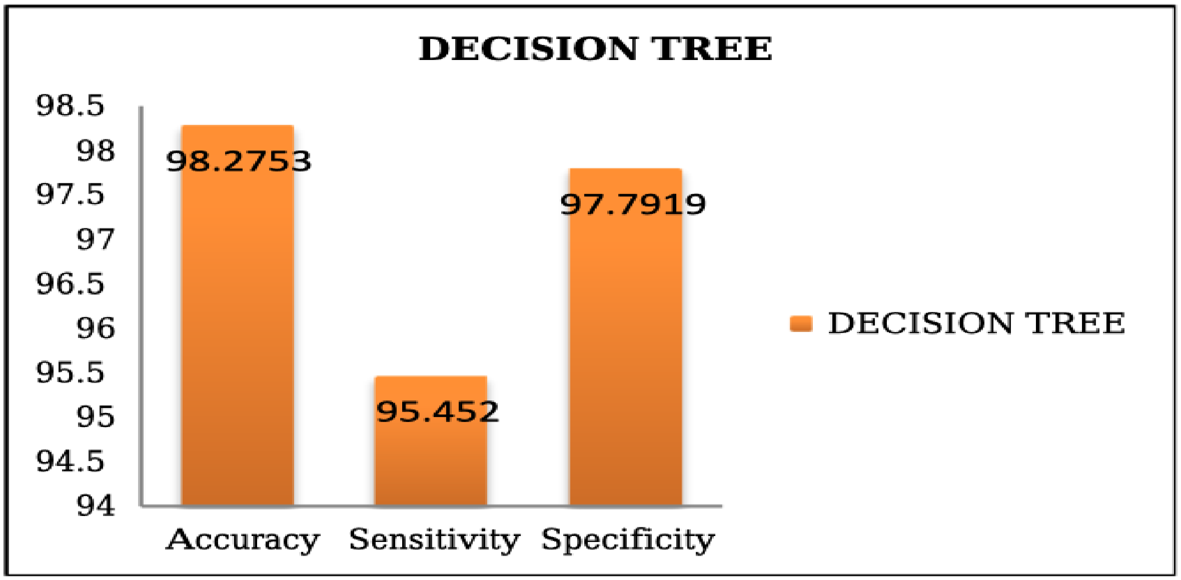
6.3 DECISION TREE TO DECISION RULES :

A decision tree can easily be transformed to a set of rules by mapping from the root node to the leaf nodes one by one.

6.4 DECISION TREE PERFORMANCE ANALYSIS :

METHOD	**DECISION TREE**
ACCURACY	98.2753
SENSITIVITY	95.452
SPECIFICITY	97.7919

7.GRAPHICAL REPRESENTATION



SUMMARY

The constructing decision tree techniques are generally computationally inexpensive, making it possible to quickly construct models even when the training set size is very large. Furthermore, once a decision tree has been built, classifying a test record is extremely fast.

8.RESULT ANALYSIS

The dataset consists of a total 573 records in the Heart disease database. The total records are divided into two data sets one is used for training consisting of 303 records & another for testing consists of 270 records. The data mining tool MATLAB is used for experiments. Initially the dataset contained some fields, in which some value in the records was missing. These were identified and replaced with most appropriate values using the Replace Missing Values filter from MATLAB. The ReplaceMissingValues filter scans all records & replaces missing values with mean mode method. This process is known as Data Pre-processing. After preprocessing the data, data mining classification techniques such as Neural Networks, Decision Trees, & Naive Bayes were applied. A confusion matrix is obtained to calculate the accuracy of classification. A confusion matrix shows how many instances have been assigned to each class. In our experiment we have two classes, and therefore we have a 2x2 confusion matrix.

Class a = YES (has heart disease)

Class b = NO (no heart disease)

9.CONFUSION MATRIX

	a(has heart disease)	b(no heart disease)
a(has heart disease)	TP	FN
b(no heart disease)	FP	TN

TP (True Positive): It denotes the number of records classified as true while they were actually true.

FN (False Negative): It denotes the number of records classified as false while they were actually true.

FP (False Positive): It denotes the number of records classified as true while they were actually false.

TN (True Negative): It denotes the number of records classified as false while they were actually false.

Confusion matrix obtained for three classification methods with 13 attributes

CONFUSION MATRIX FOR DECISION TREES

	a	b
a	123	4
b	5	138

The classification task is to generalize well on unseen/independent data. A classifier is

learned on training/learning data and then tested on data that has not been used for learning (unseen test data). There exist many measures to assess performance of a classifier and a lot of techniques to create training and test data in order to estimate generalization ability of a classifier on test (unseen) data.

Heart disease dataset: UCI Machine Learning Repository.

10.CHARACTERISTICS OF A DATA SET

Data set characteristics	Multivariate
Attribute characteristics	Real
Associated taks	Classification
Number of Instances	573
Number of Attributes	13

CLASS INFORMATION

The PHR pattern classification for the three classes are.

– Category I   (Normal)
– Category II (Disease)

11.CONCLUSION

The overall objective of our work is to predict more accurately the presence of heart disease. In this paper, the UCI repository dataset is used to get more accurate results. Three data mining classification techniques were applied namely Decision trees and Naive Bayes. From results, it has been seen that Decision trees provide accurate results as compared to Naive Bayes. This system can be further expanded. It can use more number of inputs. Other data mining techniques can also be used for prediction e.g. Clustering, Time series, Association rules. The text mining can be used to mine huge amounts of unstructured data available in healthcare industry databases.

REFERENCES

Author : A. Sankari Karthiga

Title of Paper : Early Prediction of Heart Disease Using decision tree algorithm

Publisher Name : International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)

Year of Publication : March, 2017.

URL:
https://www.researchgate.net/publication/315023624_Early_Prediction_of_Heart_Disease_Using_Decision_Tree_Algorithm

1) ShantakumarB.Patil, Y.S.Kumaraswamy “Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network”. ISSN 1450-216X Vol.31 No.4 (2009), pp.642-656.

URL :
https://www.researchgate.net/publication/254938414_A_Data_Mining_Approach_for_Prediction_of_Heart_Disease_Using_Neural_Networks

2)SellappanPalaniappan, RafiahAwang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.8,August 2008

URL :

https://www.researchgate.net/publication/4329399_Intelligent_heart_disease_prediction_system_using_data_mining_techniques