ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

Novel Approach for Heart Disease Prediction Using Decision Tree Algorithm

Gadoya Komal[1], DR.Vipul Vekariya[2]

[1]PG Student, Dept. of CE, Noble Group of Institute, Junagadh, Gujrat, India

[2]Assistant Professor, Dept. of CE, Noble Group of Institute, Junagadh, Gujrat, India

ABSTRACT:The procedure of applying intelligent as well as perceptive methods for extracting data patterns is called Data mining. There are so many techniques generally used to identify information and in decision making for knowledge presentation. Extraction of data in these way that they are useful in areas like decision making as well as for valuable forecasting also in computation and predictions. The Healthcare, clinical and medical fields are rich in information but are not exactly used to its area. The healthcare industry collects huge amounts of healthcare information that are not "mined" means extracted to discover hidden information. For effective decision making, healthcare organizations are faced with challenges to provide cost-effective, efficient as well as richer quality of patient care. In short, to discover the relations which present between connect parameters in a database is the subject of data mining. This research has developed a prototype Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques, namely, Decision Trees i.e. ID3.By using medical profiles of patients such as age, gender, blood pressure and blood sugar, chest pain, ECG graph etc, it can predict the likelihood of patients getting a heart disease or not. This proposed system is implemented in MATLAB as an application that takes parameters of medical test as an input. It shows as a training tool to train nurses ,medical students, and also for fresher in medical analysis to diagnose patients with heart disease.

KEYWORDS: Data mining, Gini Index, Information Gain, Gain ratio, heart disease, Decision tree algorithm, decision support

I. INTRODUCTION

Data mining means to extracting or "mining" knowledge from large amounts of data available in sources.It is the technique of identifying potentially useful, novel, true, most important understandable pattern in data by using of databases. Hence it is called necessary method of Knowledge Discovery. It is one type of conversion of data into knowledge is useful for decision making called to as data mining. Finding for the association, relationships also global patterns from large size of databases called it as non trivial process of Data mining. Giudici introduced the data mining as "a process of selection, exploration and modelling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of database" [5].This Knowledge Discovery process builds of an step by step sequential tasks of data pre-processing like data cleaning, integration, data selection, identification of patterns and finally presentation of knowledge.

Some hospital information system are today adopted by hospitals for handling as well managing their healthcare or patient as well clinical data [2]. Decision support systems already used by some hospitals already used, but are small in amount. Some hospitals have decision support systems that limited in size. Decisions in the hospital are generally done with respect to experience and guidance of the doctors rather than on the richer, non trivial, useful, mined knowledge that hidden in the database records. This may causes to errors, mistakes unwanted biases also leads to excessive medical costs that affects the patients service quality. WHO estimated by 2030, there are 23.6 million people will die cause of Heart disease as written in [15]. Prediction by using data mining techniques generates accurate result of disease. Hence, the proposed system used to integrate as well as generates proper decision in the clinic. Here patient records generates

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

in computer-based. This system can improve overall patient safety, reduce medical decision mistake and reduces undesirable practice variation, and magnify patient outcome. This application is promising as data modelling and analysis tools, say as for generate a knowledge rich environment that can help to significantly do improvement the quality of clinical decisions that leads to make the system accurate and highly performable provided as a strategy by data mining.

II. PROBLEM STATEMENT

Hospitals today adopted some hospital information system for handling as well managing their healthcare or patient as well clinical data. Decision support systems already used by some hospitals already used, but are small in amount. It can answer queries which are in simple forms like “What is the average age in years of patients who have heart disease with high BP rate?” and “How many of heart disease surgeries had resulted in hospital stays longer than 35 days?”, “Number of percentage of the female patients who are single, having higher heart rate below 26 years old, and who have been treated for cancer.” But, they cannot answer complex type of queries like “Clarify the required preoperative predictors that increase the length of hospital stay”, “Specify patient data on cancer, should treatment include radiation only, chemotherapy alone or both radiation and chemotherapy?”,“Given patient records, predict the probability of patients that having higher blood pressure and having high heart rate getting a heart disease.” Decisions in the hospital are generally done with respect to experience and guidance of the doctors rather than on the richer, non trivial, useful, mined knowledge that hidden in the database records. This may causes to errors, mistakes unwanted biases, and excessive medical expenditure which affects the quality of service provided to patients. Wu, et al proposed that For reduction of this medical errors, improvement in patient safety, enhance patient outcome and decrease unwanted practice variation could be improve by merging criteria of clinical decision support and computer-based patient records [7].

III. RESEARCH OBJECTIVES

The medical/clinical database is a large database that maintains various medical data types such as patient data history, treatment records, medication profiles, radiology as well as pathology report, signal moreover images. For management of healthcare or patient data, most hospitals today use hospital information systems. But, that produces more sized of data, out of them some amount of data are useful, and others are not. There is a presence of some unknown information in these data which largely not achieved. So there is need to convert this data into useful information which can useful for healthcare systems and fresher practitioners for making proper clinical decisions. The main objective criteria of this proposed research work is to develop a Decision Support system in Heart Disease Prediction System (HDPS) using data mining method namely, decision tree(ID3). HDPS is implemented as an application in matlab which can responds user queries, this application can discover and extract hidden knowledge such as relationships and patterns with respect to heart disease from historical database of heart disease [9]. This proposed system gives the report of the patient that states whether that considered patient having the heart disease or not. This application is promising as data modelling and analysis tools, say as for generate a knowledge rich environment that can help to significantly do improvement the quality of clinical decisions that leads to make the system accurate and highly performable provided as a strategy by data mining [16].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

Clinical databases have a giant quantity of information about patients say patient profile and their medical conditions. The term cardiovascular disease signifies the diverse disease that causes the heart [18]. Major reason of casualties in the world is the Heart disease. Cardio vascular disease kills one person each 34 seconds within the US. Coronary heart disease, Cardiomyopathy and some categories of heart diseases are cardiovascular diseases. This term “cardiovascular disease” includes a wide range of conditions which affect the heart and the blood vessels and the manner in which blood is pumped and circulated through the body. This Cardiovascular disease (CVD) generates results in severe illness, disability, and death [19]. Here, 13 attributes has been used for consistency [17].

The attribute “Diagnosis” is known as the predictable attribute with value “1” for patients with heart disease and value “0” for patients with no heart disease. “Patient’s test” is employed as a record, last attribute as output and, the remaining are input attributes.

No	Name	Description
1	Age	Age in years
2	gender	1 = male 0 = female
3	Cp	Chest pain type: A = typical angina B = atypical angina C= non-anginal pain D =asymptomatic
4	Trestbps	Resting blood pressure (in mm Hg)
5	Chol	Serum cholesterol in mg/dl
6	Fbs	Fasting blood sugar > 120mg/dl: 1 = true 0 = false
7	Restecg	Resting electrocardiographic results: A= normal B= having ST-T wave abnormality C=showing probable or left ventricular hypertrophy by Estes’criteria
8	Thalach	Maximum heart rate Achieved
9	Exang	Exercise induced angina: 1 = yes 0 = no
10	Old peak ST	Depression induced by exercise relative to rest
11	Slope	The slope of the peak exercise segment : A = up sloping B = flat C= down sloping

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

12	Ca	Number of major vessels colored by fluoroscopy that ranged between 0 and 3.
13	Diagnosis	Diagnosis classes: 0 = healthy 1= patient who is subject to possible heart disease
14	Thal	X = normal Y= fixed defect Z= reversible defect

IV. RELATED WORK

Many researchers have put their attention on Heart Disease Prediction by using many Data Mining techniques. Some of them are as follows.

1. In [3] authors represent the prediction by using naive bayes, decision tree and neural network techniques .It reduces treatment costs. But The size of the dataset used in this research is still quite small.
2. In [6] authors used Short Text Classification, Smoothing, Naive Bayes.Here, Smoothing methods are able to greatly improve the accuracy of Naive Bayes for short text classification but they can only slightly help for normal documents.
3. In [1] Naïve Bayes and Jelinek- Smercer smoothing is used. It provides better performance and system more accurate but it uses Naïve Bayes Classifier technique is mainly applicable when the dimensionality of the inputs is high.
4. In [8] authors used data mining method of Naïve Bayes. Here, ease of model interpretation and access to detailed information and accuracy but here,Just categorical data used.
5. In [10] Proposed algorithm uses Naive Bayesian Classification which overcome unwanted biases, errors which affects the quality of service. But here, challenge would be to integrate data mining and text mining.
6. In [11] Artificial Neural Networks (ANN) used for improving performance but generates huge amounts of complex data.

V. PROPOSED ALGORITHM

Basically, the learning of the decision tree from class labeled training tuples is called Decision tree induction. We can define, decision tree is like a flow chart tree structure, here internal node also called non leaf node states a test on the attribute, each branch here states an outcome of the test, each leaf node called as terminal node denotes a class label and in this tree structure topmost node denotes the root node.

Decision Tree learning is one of the most widely used and practical methods for inductive inference over supervised data. A decision tree represents a procedure or technique for classifying categorical data based on their attributes. It is also efficient for processing large amount of data, so is often used in data mining application [14]. There is not necessary any domain knowledge or any type of parameter setting for the decision tree construction and therefore exact way for knowledge discovery. Decision tree's representation of extracted knowledge in tree intuitive form and easy to understand by humans.

Decision Tree Algorithm – ID3

During the late 1970s and early 1980s, J. Ross Quinlan, which was a researcher in machine learning, founded a decision tree algorithm also known as ID3 (Iterative Dichotomiser).

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

- Decide which attribute (splitting-point) to test at node ***K*** by determining the "best" way to separate or partition the tuples in ***DP*** into individual classes.
- The splitting criteria is determined so that, ideally, the resulting partitions at each branch are as "pure" as possible.
- A partition is pure if all of the tuples in it belong to the same class.

<u>Algorithm:</u>
Generate decision tree. Generate a decision tree from the training tuples of data partition *DP*.

Input:

- Data partition, *DP*, that is a set of training tuples with their associated class labels;
- *attribute_set*, which is the set of candidate attributes;
- *attribute_selection_process*, defines a process to determine the splitting criterion which "best" partitions the data tuples into individual classes. It is consist of a *splitting attribute* and which is either a *split point* or *splitting subset*.

Output: A decision tree.

Method:

1) create a node *K*;
2) if tuples in *DP* are all of the same class, *C* then
3) return *K* as a leaf node labeled with the class *C*;
4) if *attribute_set* is empty then
5) return *K* as a leaf node labeled with the majority class in *DP*; // majority voting
6) apply *attribute_selection_process* (*DP*, *attribute_set*) to find the "best" *splitting criterion*;
7) label node *K* with *splitting criterion*;
8) if *splitting attribute* is discrete-valued and multiway splits allowed then // not restricted to binary trees
9) *attribute_set*←
    attribute_set-splitting_attribute;
   // remove *splitting attribute*
10) for each outcome *j* of *splitting criterion* // partition the tuples and grow sub trees for each partition
11) let *DPj* be the set of data tuples in *DP* satisfying outcome *j*; // a partition
12) if *DPj* is empty then
13) attach a leaf labeled with the majority class in *DP* to node *K*;
14) else attach the node returned by Generate_decision_tree
15) (*DPj*, *attribute_set*) to node *K*;
    Endfor;
16) return *K*;

This splitting partitioning stops when any one of the following conditions occurs a true:

1. If for all of the tuples in partition *DP* known at node *K* , that belong to the same class (steps 2 and 3), or
2. If there are no attributes on which the tuples may be partitioned further as shown in step 4. In this case, majority voting is done as shown in step 5. This involves converting node *K* into a leaf and labeling it with the most common class in *DP*. Recursively, the node tuple's class distribution may be stored.
3. If there are no tuples for that given branch, means that a partition *DPj* is empty (step 12). In that case, a leaf is generated with the majority class in *DP* (step 13).

- **Information Gain**

The entropy means Information Gain approach generally selects the actual splitting attribute which minimizes the value of entropy and thus maximising the Information Gain for identification of the splitting attribute of the Decision Tree.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

There is must be calculate the Information Gain for each of the attribute after that selection of the attribute that maximizes the Information Gain is done. The Information Gain for each attribute can be calculated using the given below formula which states in Han and Kamber 2006; Bramer 2007:

$$E = \sum_{i=1}^{m} pi \log_2 pi \qquad \text{eq. (1)}$$

Here m states for the no. of classes of the target attributes and here, Pi states for the no. of occurrences of class i divided by the total number of instances means states for probability of i occurring.

- **Gini Index**

Measurement of the impurity of the data counted by the Gini index. The Gini Index is calculated for each attribute in the data set. If there are m classes of the target attribute, with the probability of the ith class being Pi, the Gini Index is defined as (Bramer 2007):

$$\text{Gini Index} = 1\text{-}\sum_{i=1}^{m} pi^2 \qquad \text{eq. (2)}$$

The splitting attribute referred as the attribute has the largest reduction in the counted value of the Gini Index.

- **Gain Ratio**

For reduction of the effect of the biasing that resulting from the general use of Information Gain, another alternative known as Gain Ratio was given by the Ross Quinlan in 2007.The Information Gain gives biased of tests with showing too many outcomes. Generally, it used to select attributes that having a many number of values.

$$\text{Gain Ratio} = \text{Information Gain / Split Info} \qquad \text{eq. (3)}$$

Here, the split information is a one type of value that depends on the sums of that frequency.

- **Pruning**

Reduced error pruning is used for pruning the decision rules that are extracted previously.It is the fastest pruning technique. It is used to generate accurate as well as small decision rules. Reduced error pruning generates compact decision rules. Hence, it generates the number of extracted rules.

- **Steps to implement Decision tree for patient records:**

- Enter data of patient record.
- There are two classes in which we have to classify the data are.
    - 0:HD absent ,
    - 1:HD present
- Decide the probability of each attribute for both the classes using the database with result as training.
- Apply decision tree and identify the impact as well as relationship that present between medical attributes with relation to the predictable state of heart disease.
- Attribute measurement by using Information Gain, Gain Ratio and Gini index and tree pruning.
- Decide the class for patient record.

- **Flowchart for implementation of classification on patient data**

Here, FIG 1 shows the flow of the algorithm step by step. In the last step in says about the risk means probability of patients to having heart diseases or not.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

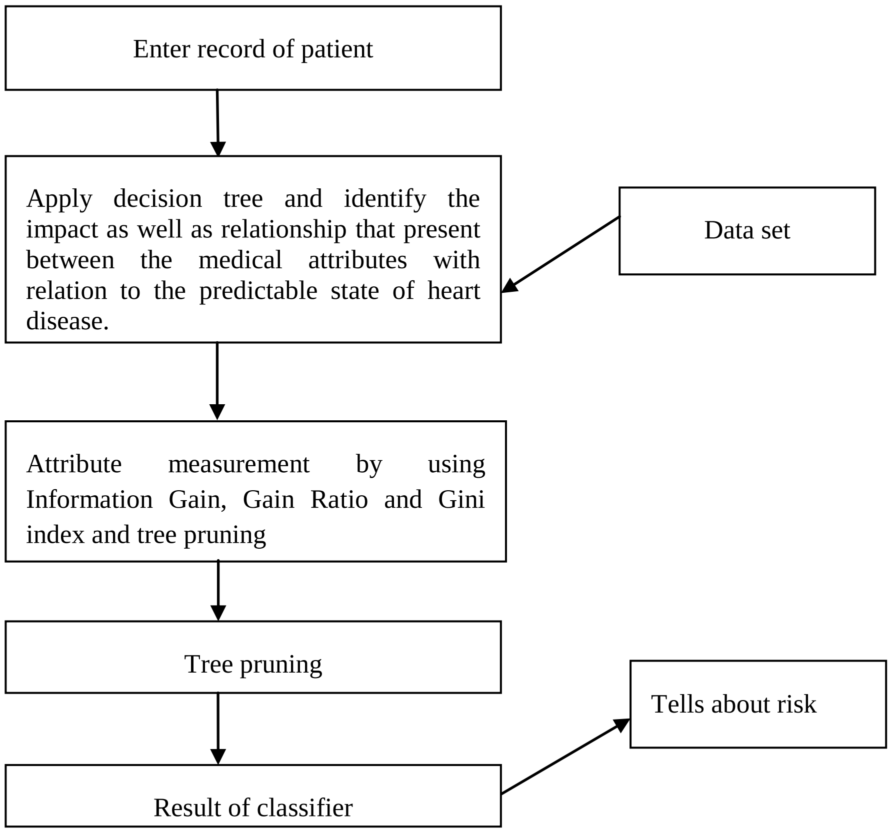

Fig.1 Implementation of ID3 (decision tree) algorithm on the patient data.

VI. PERFORMANCE ANALYSIS

Negi are here states the negative tuples which were correctly labelled by the classifier. False positives means False_Posi states the negative tuples which were incorrectly labelled by the classifier and false negatives states positive tuples which were incorrectly labelled by the classifier.
Recognition rate called as a Sensitivity or true positive rate. Here, specificity and sensitivity measures can be using to calculate performance. Precision defines percentage of samples that labelled with Yes. There is given formula of calculating Sensitivity.

$$\text{Sensitivity} = \text{True\_Posi} / \text{Posi} \qquad \text{eq. (4)}$$

Specificity called as a true negative rate. True_Posi states for the number of true positives means "Present" samples which were correctly classified. Posi states for the number of positive samples. There is given formula of calculating Specificity.

$$\text{Specificity} = \text{True\_Negi} / \text{Negi} \qquad \text{eq. (5)}$$

True_Negi states for the number of true negatives means Absent samples which were correctly classified. Negi states for the number of negative samples. False_Posi states for the number of false positives means Absent samples which were incorrectly labelled with Yes.

$$\text{Precision} = \textit{True\_Posi} / (\textit{True\_Posi} + \textit{False\_Posi}) \qquad \text{eq. (6)}$$

$$\text{Accuracy} = \text{Sensitivity}\ [\text{Posi} \backslash (\text{Posi} + \text{Negi})] + \text{Specificity}\ [\text{Negi} \backslash (\text{Negi} + \text{Posi})] \qquad \text{eq. (7)}$$

The true negatives, true positives, false negatives and false positives are also useful for assessing the benefits as well as cost associated with a classification model.

VII. CONCLUSION AND FUTURE SCOPE

Development of Decision Support in Heart Disease Prediction can be done by using decision tree algorithm. This system with the use of data mining, extracts hidden and useful knowledge from a historical database of the heart disease. This model also could gives answer of complex queries, and an ease of access to historic information and easy to model interpretation and performed with better accuracy. This system is says to be expandable means that if more number of records or attributes can be added and it can be incorporated as well and also generation of new significant decision rules can be done underlying Data Mining technique. This Decision Trees give the result, that are easy to read and easy to interpret by humans. This feature to read detailed profiles of patients is only available in Decision Trees. Presently the system has been using 13 attributes of medical diagnosis. It can also use other data mining techniques and additional attributes for prediction. Though, various classification techniques are widely used for Disease Prediction, Decision Tree classifier is selected for its simplicity and accuracy. Different attribute selection measures like Information Gain, Gain Ratio, Gini Index and Distance measure can be used for better accuracy.

REFERENCES

1. Ms.Rupali R.Patil **,**“Heart Disease Prediction System using Naive Bayes and Jelinek-mercer smoothing” ,Vol. 3, Issue 5, May 2014 © IJARCCE
2. Obenshain, M.K: “Application of Data Mining Techniques to Healthcare Data”, Infection Control and Hospital Epidemiology, 25(8), 690–695, 2004.
3. Sellappan Palaniappan, Rafiah Awang, Intelligent Heart Disease Prediction System Using Data Mining Techniques, 978-1-4244-1968-5/08/$25.00 ©2008 IEEE.
4. Han, J., Kamber, M.: “Data Mining Concepts and Techniques”, Morgan Kaufmann Publishers, 2006.
5. Giudici, P.: “Applied Data Mining: Statistical Methods for Business and Industry”, New York: John Wiley, 2003.
6. Eman AbuKhousa, Piers Campbell-“Predictive Data Mining to Support Clinical Decisions:An Overview of Heart Disease Prediction Systems”, 978-1-4673-1101-4/12/$31.00 ©2012 IEEE
7. Wu, R., Peters, W., Morgan, M.W.: “The Next Generation Clinical Decision Support: Linking Evidence to Best Practice”, Journal Healthcare Information Management. 16(4), 50-55, 2002.
8. Mrs.G.Subbalakshmi , Mr. K. Ramesh, Mr. M. Chinna Rao,- “Decision Support in Heart Disease Prediction System using Naive Bayes”, ISSN : 0976-5166Vol. 2 No. 2 Apr-May 2011.
9. “Heart diseases” from http://en.wikipedia.org/wiki/Heart_disease
10. Shadab Adam Pattekari and Asma Parveen, “PREDICTION SYSTEM FOR HEART DISEASE USING NAIVE BAYES”, *International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624. Vol 3, Issue 3, 2012, pp 290-294.*
11. A.Sudha, P.Gayathri, N.Jaisankar - ”Effective Analysis and Predictive Model of Stroke Disease using Classification Methods”,*International Journal of Computer Applications (0975 – 8887) Volume 43– No.14, April 2012 .*
12. Mohammad Taha Khan, Dr. Shamimul Qamar and Laurent F. Massin, A Prototype of Cancer/Heart Disease Prediction Model Using Data Mining, International Journal of Applied Engineering Research, 2012.
13. Ms. Ishtake S.H, Prof. Sanap S.A., “Intelligent Heart Disease Prediction System Using Data Mining Techniques”, International J. of Healthcare & Biomedical Research, 2013.
14. Nidhi Bhatla, Kiran Jyoti, An Analysis of Heart Disease Prediction using Different Data Mining Techniques, International Journal of Engineering Research & Technology (IJERT), 2012.
15. Miss. Chaitrali S. Dangare, Dr. Mrs. Sulabha S. Apte, A data mining approach for prediction of heart disease using neural networks, international journal of computer engineering and technology, 2012.
16. Beant Kaur , Williamjeet Singh ,”Review on Heart Disease Prediction System using Data Mining Techniques” International Journal on Recent and Innovation Trends in Computing and Communication(IJRITCC),October 2014.
17. 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.
18. Clinical Data Mining: a Review J. Iavindrasana, G. Cohen, A. Depeursinge, H. Müller, R. Meyer, A.Geissbuhler University and Hospitals of Geneva, Switzerland.
19. Shanthi Mendis; Pekka Puska; Bo Norrving; World Health Organization (2011). *Global Atlas on Cardiovascular Disease Prevention and Control*. World Health Organization in collaboration with the World Heart Federation and the World Stroke Organization. pp. 3–18. ISBN 978-92-4-156437-3.
20. Mary K. Obenshain , MAT ; “Application of Data Mining Techniques to Healthcare Data,” *Infection Control and Hospital Epidemiology ,* Vol. 25, No. 8, pp. 690-695, August 2004.