ISSN (ONLINE):2395-695X
ISSN (PRINT):2395-695X

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol.3, Issue.3, March 2017

Early Prediction of Heart Disease Using Decision Tree Algorithm

A. Sankari Karthiga[1], M. Safish Mary[2], M. Yogasini[3]
M.Phil. Scholar[1], Assistant Professor[2], Assistant Professor[3],
Mother Teresa Women's University, Kodaikanal[1]
St. Xavier's College, Tirunelveli[2]
Sadakathdulla Appa College, Tirunelveli[3]
sankarikarthiga.sk@gmail.com

***Abstract*-For processing of large amount 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 field due to its application in healthcare domain. Classification and prediction of medical datasets poses challenges in Medical Data Mining. The 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 and Naïve Bayes are applied in this research for predicting heart attacks. The research result shows prediction accuracy of 99%. Data mining enable the health sector to predict patterns in the dataset.**

***Index Terms*- Decision Tree Algorithm, Naïve Bayes Algorithm.**

I. INTRODUCTION

1.1. DATA MINING

Data Mining is about explaining the past and predicting the future by means of data analysis. Data mining is a multi-disciplinary field that combines statistics, machine learning, artificial intelligence and database technology. The value of data mining applications is often estimated to be very high. Many businesses have stored large amounts of data over years of operation, and data mining is able to extract very valuable knowledge from this data. The businesses are then able to leverage the extracted knowledge into more clients, more sales, and greater profits. This is also true in the engineering and medical fields.

1.1.1. Statistics

The science of statistics is to collecting, classifying, summarizing, organizing, analyzing, and interpreting data.

1.1.2. Artificial Intelligence

The study of computer algorithms is to dealing with the simulation of intelligent behaviour in order to perform those activities that are normally thought to require intelligence.

1.1.3. Machine Learning

The study of the computer algorithms aim is to learn in order to improve automatically through experience.

1.1.4. Database

The science and technology of collecting, storing and managing data so users can retrieve, add, update or remove such data.

1.1.5. Data warehousing

The science and technology of collecting, storing and managing data with advanced multi-dimensional reporting services in support of the decision-making processes.

1.1.6. Explaining the Past

Data mining explains the past through data exploration.

1.1.7. Predicting the Future

Data mining predicts the future by means of modeling.

1.1.8. Data Exploration

Data Exploration is about describing the data by means of statistical and visualization techniques. We explore data in order to bring important aspects of that data into focus for further analysis.

"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 healthcare industry. Nowadays healthcare industry generates 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 are unacceptable.

According to survey of WHO, 17 million total global deaths are due to heart attacks and strokes. The deaths due to heart disease in many countries occur due to work overload, mental stress and many other problems. Overall, it is found as primary reason behind death in adults. Diagnosis is complicated and important task that needs to be executed accurately and efficiently. The diagnosis is often made, based on doctor's experience & knowledge. This leads to unwanted results & excessive medical costs of treatments provided to patients.

Therefore, an automatic medical diagnosis system is designed that take advantage of collected database and decision support system. This system can help in diagnosing disease with less medical tests & effective treatments.

ISSN (ONLINE):2395-695X
ISSN (PRINT):2395-695X

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol.3, Issue.3, March 2017

1.2. MEDICAL DATA MINING

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.

The World Health Organization has estimated that 12 million deaths occurs worldwide, every year due to the Heart diseases. Half the deaths in the United States and other developed countries occur due to cardio vascular 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. Heart disease kills one person every 34 seconds in the United States. 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.

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. Regrettably all doctors do not possess expertise in every sub specialty and moreover there is a shortage of resource persons at certain places. Therefore, an automatic medical diagnosis system would probably be exceedingly beneficial by bringing all of them together. 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 have 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.

Naïve Bayes is a statistical classifier which assumes no dependency between attributes. It attempts to maximize the posterior probability in determining the class. The advantage of using naive bayes is that one can work with the naive Bayes model without using any Bayesian methods. Naive Bayes classifiers have works well in many complex real-world situations

ISSN (ONLINE):2395-695X
ISSN (PRINT):2395-695X

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol.3, Issue.3, March 2017

1.3. Heart Disease

The heart is important organ of human body part. It is nothing more than a pump, which pumps blood through the body. If circulation of blood in body is inefficient the organs like brain suffer and if heart stops working altogether, death occurs within minutes. Life is completely dependent on efficient working of the heart. The term Heart disease refers to disease of heart & blood vessel system within it.

A number of factors have been shown that increases the risk of Heart disease:

- Family history
- Smoking
- Poor diet
- High blood pressure
- High blood cholesterol
- Obesity
- Physical inactivity
- Hyper tension

Factors like these are used to analyze the Heart disease. In many cases, diagnosis is generally based on patient's current test results & doctor's experience. Thus the diagnosis is a complex task that requires much experience & high skill.

Heart disease is a broad term that includes all types of diseases affecting different components of the heart. Heart means 'cardio.' Therefore, all heart diseases belong to the category of cardiovascular diseases. Some types of Heart diseases are

1. Coronary heart disease It also known as coronary artery disease (CAD), it is the most common type of heart disease across the world. It is a condition in which plaque deposits block the coronary blood vessels leading to a reduced supply of blood and oxygen to the heart.
2. Angina pectoris it is a medical term for chest pain that occurs due to insufficient supply of blood to the heart. Also known as angina, it is a warning signal for heart attack. The chest pain is at intervals ranging for few seconds or minutes.
3. Congestive heart failure it is a condition where the heart cannot pump enough blood to the rest of the body. It is commonly known as heart failure.
4. Cardiomyopathy, it is the weakening of the heart muscle or a change in the structure of the muscle due to inadequate heart pumping. Some of the common causes of cardiomyopathy are hypertension, alcohol consumption, viral infections, and genetic defects.
5. Congenital heart disease, it also known as congenital heart defect, it refers to the formation of an abnormal heart due to a defect in the structure of the heart or its functioning. It is also a type of congenital disease that children are born with.

ISSN (ONLINE):2395-695X
ISSN (PRINT):2395-695X

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol.3, Issue.3, March 2017

6. Arrhythmias it is associated with a disorder in the rhythmic movement of the heartbeat. The heartbeat can be slow, fast, or irregular. These abnormal heartbeats are caused by a short circuit in the heart's electrical system.
7. Myocarditis it is an inflammation of the heart muscle usually caused by viral, fungal, and bacterial infections affecting the heart. It is an uncommon disease with few symptoms like joins pain, leg swelling or fever that cannot be directly related to the heart.

1.4. 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.

1.5. Naive Bayes

Naive Bayes classifier is based on Bayes theorem. This classifier algorithm uses conditional independence, means it assumes that an attribute value on a given class is independent of the values of other attributes.

1.6. 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.

II. Classification Using Decision Tree Algorithm

2.1. Introduction

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.

ISSN (ONLINE):2395-695X
ISSN (PRINT):2395-695X

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol.3, Issue.3, March 2017

2.2. Algorithm

The core algorithm for building decision trees called C**4.5** by J. R. Quinlan which employs a top-down, greedy search through the space of possible branches with no backtracking. C4.5 uses *Entropy* and *Information Gain* to construct a decision tree.

2.3. 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). 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 an 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:

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

2.4. 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 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 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.

2.5. 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.

ISSN (ONLINE):2395-695X
ISSN (PRINT):2395-695X

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol.3, Issue.3, March 2017

III. CLASSIFICATION USING NAIVE BAYES CLASSIFIER

A. INTRODUCTION

The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.

B. ALGORITHM

Bayes theorem provides a way of calculating the posterior probability, P (c|x), from P(c), P(x), and P (x|c). Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.

Likelihood — Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

Posterior Probability — Predictor Prior Probability

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

- *P (c|x)* is the posterior probability of class (target) given predictor (attribute).
- *P(c)* is the prior probability of class.
- *P (x|c)* is the likelihood which is the probability of predictor given class.
- *P(x)* is the prior probability of predictor.

Thus, we can write:

$$\text{Prior probability for GREEN} \propto \frac{\text{Number of GREEN objects}}{\text{Total number of objects}}$$

$$\text{Prior probability for RED} \propto \frac{\text{Number of RED objects}}{\text{Total number of objects}}$$

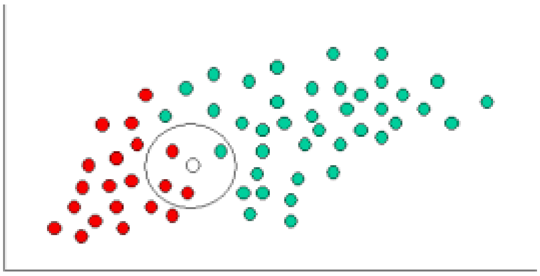
Since there is a total of 60 objects, 40 of which are GREEN and 20 RED, our prior probabilities for class membership are:

$$\text{Prior probability for GREEN} \propto \frac{40}{60}$$

$$\text{Prior probability for RED} \propto \frac{20}{60}$$

ISSN (ONLINE):2395-695X
ISSN (PRINT):2395-695X

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol.3, Issue.3, March 2017



Having formulated our prior probability, we are now ready to classify a new object (WHITE circle). Since the objects are well clustered, it is reasonable to assume that the more GREEN (or RED) objects in the vicinity of X, the more likely that the new cases belong to that particular color. To measure this likelihood, we draw a circle around X which encompasses a number (to be chosen a priori) of points irrespective of their class labels. Then we calculate the number of points in the circle belonging to each class label. From this we calculate the likelihood:

$$\text{Likelihood of } X \text{ given GREEN} \propto \frac{\text{Number of GREEN in the vicinity of } X}{\text{Total number of GREEN cases}}$$

$$\text{Likelihood of } X \text{ given RED} \propto \frac{\text{Number of RED in the vicinity of } X}{\text{Total number of RED cases}}$$

From the illustration above, it is clear that Likelihood of X given GREEN is smaller than Likelihood of X given RED, since the circle encompasses 1 GREEN object and 3 RED ones. Thus:

$$\text{Probability of } X \text{ given GREEN} \propto \frac{1}{40}$$

$$\text{Probability of } X \text{ given RED} \propto \frac{3}{20}$$

Although the prior probabilities indicate that X may belong to GREEN (given that there are twice as many GREEN compared to RED) the likelihood indicates otherwise; that the class membership of X is RED (given that there are more RED objects in the vicinity of X than GREEN). In the Bayesian analysis, the final classification is produced by combining both sources of information, i.e., the prior and the likelihood, to form a posterior probability using the so-called Bayes' rule (named after Rev. Thomas Bayes 1702-1761).

$$\text{Posterior probability of } X \text{ being GREEN} \propto$$
$$\text{Prior probability of GREEN} \times \text{Likelihood of } X \text{ given GREEN}$$
$$= \frac{4}{6} \times \frac{1}{40} = \frac{1}{60}$$
$$\text{Posterior probability of } X \text{ being RED} \propto$$
$$\text{Prior probability of RED} \times \text{Likelihood of } X \text{ given RED}$$
$$= \frac{2}{6} \times \frac{3}{20} = \frac{1}{20}$$

Finally, we classify X as RED since its class membership achieves the largest posterior probability.

Naive Bayes can be modelled in several different ways including normal, lognormal, gamma and Poisson density functions:

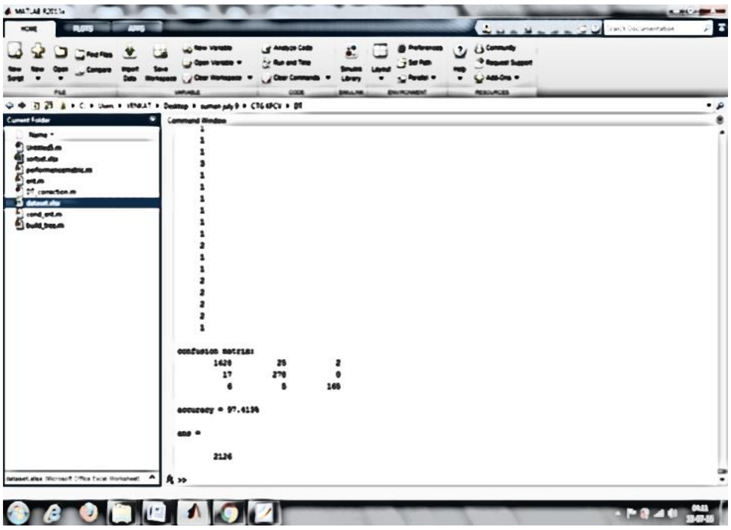
ISSN (ONLINE):2395-695X
ISSN (PRINT):2395-695X

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol.3, Issue.3, March 2017

$$p(x_k \mid C_j) = \begin{cases} \dfrac{1}{\sigma_{kj}\sqrt{2\pi}} \exp\left(\dfrac{-(x-\mu_{kj})^2}{2\sigma_{kj}^2}\right), \quad -\infty < x < \infty, -\infty < \mu_{kj} <, \sigma_{kj} > 0 & \text{Normal} \\ \mu_{kj} : \text{mean}, \ \sigma_{kj} : \text{standard deviation} & \\ \dfrac{1}{x\sigma_{kj}(2\pi)^{1/2}} \exp\left\{\dfrac{-[\log(x/m_{kj})]^2}{2\sigma_{kj}^2}\right\}, \quad 0 < x < \infty, m_{kj} > 0, \sigma_{kj} > 0 & \text{Lognormal} \\ m_{kj} : \text{scale parameter}, \ \sigma_{kj} : \text{shape parameter} & \\ \dfrac{\left(\dfrac{x}{b_{kj}}\right)^{c_{kj}-1}}{b_{kj}\Gamma(c_{kj})} \exp\left(\dfrac{-x}{b_{kj}}\right), \quad 0 \le x < \infty, b_{kj} > 0, c_{kj} > 0 & \text{Gamma} \\ b_{kj} : \text{scale parameter}, \ c_{kj} : \text{shape parameter} & \\ \dfrac{\lambda_{kj}\exp(-\lambda_{kj})}{x!}, \quad 0 \le x < \infty, \lambda_{kj} > 0, x = 0,1,2,\ldots & \text{Poisson} \\ \lambda_{kj} : \text{mean} & \end{cases}$$

C. Performance Analyses

(i) Decision Tree Classifier-Cross Validation (Experimental Results)

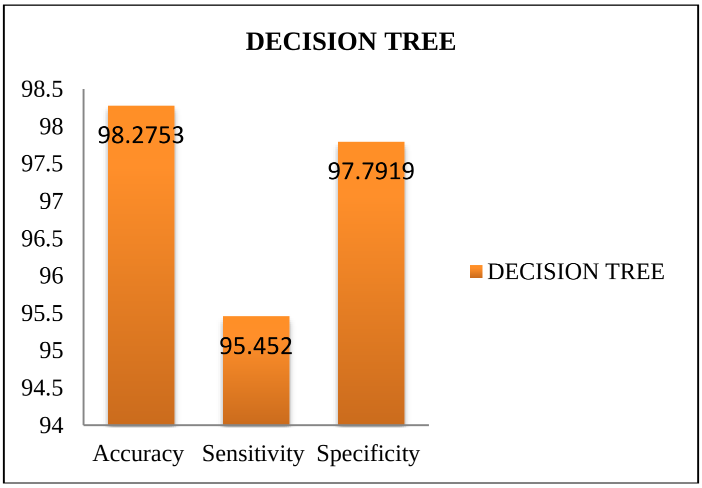


(ii) Decision Tree Performance Metrics

METHOD	DECISION TREE
Accuracy	98.2753
Sensitivity	95.452
Specificity	97.7919

ISSN (ONLINE):2395-695X
ISSN (PRINT):2395-695X

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol.3, Issue.3, March 2017

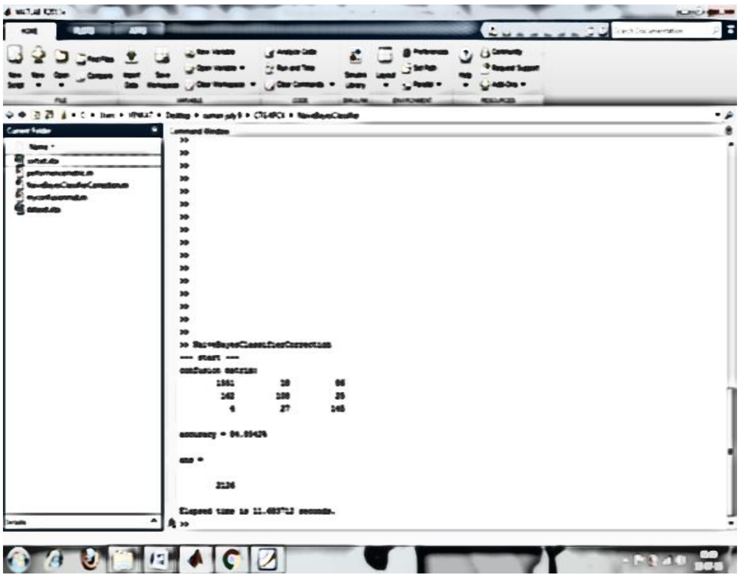


(iii) SUMMARY

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.

D. NAÏVE BAYES

(i) EXPRIMENTAL RESULTS

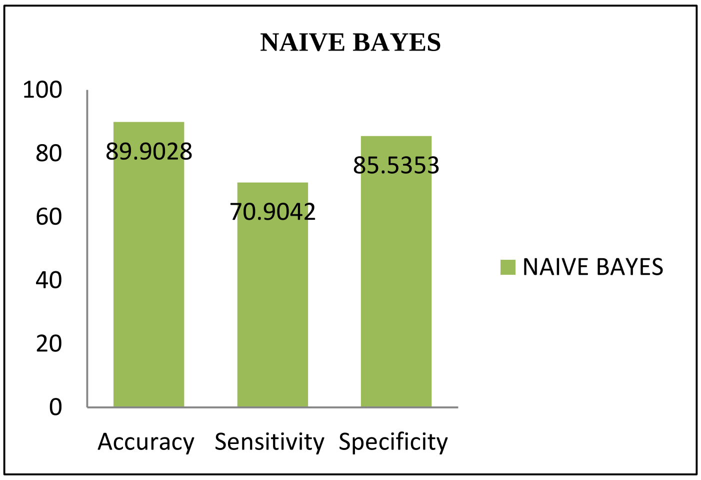


(ii) NAÏVE BAYES PERFORMANCE METRICS

METHOD	NAIVE BAYES
Accuracy	89.9028
Sensitivity	70.9042
Specificity	85.5353

ISSN (ONLINE):2395-695X
ISSN (PRINT):2395-695X

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol.3, Issue.3, March 2017



(iii) SUMMARY

Poisson variables are regarded here as continuous since they are ordinal rather than truly categorical. For categorical variables, a discrete probability is used with values of the categorical level being proportional to their conditional frequency in the training data.

IV. RESULT ANALYSIS

The dataset consists of total 573 records in Heart disease database. The total records are divided into two data sets one is used for training consists of 303 records & another for testing consists of 270 records. The data mining tool MATLAB is used for experiment.

Initially dataset contained some fields, in which some value in the records was missing. These were identified and replaced with most appropriate values using 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 pre-processing 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)

ISSN (ONLINE):2395-695X
ISSN (PRINT):2395-695X

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol.3, Issue.3, March 2017

V. 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 Naive Bayes

	a	b
a	110	5
b	10	145

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.

Characteristics Of A Data Set

Data Set Characteristics	**Multivariate**
Attribute Characteristics	Real
Associated tasks	Classification
Number of Instances	573
Number of Attributes	13

ISSN (ONLINE):2395-695X
ISSN (PRINT):2395-695X

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol.3, Issue.3, March 2017

CLASS INFORMATION:

The PHR pattern classification for the three class are.

- Category I (Normal)
- Category II (Disease)

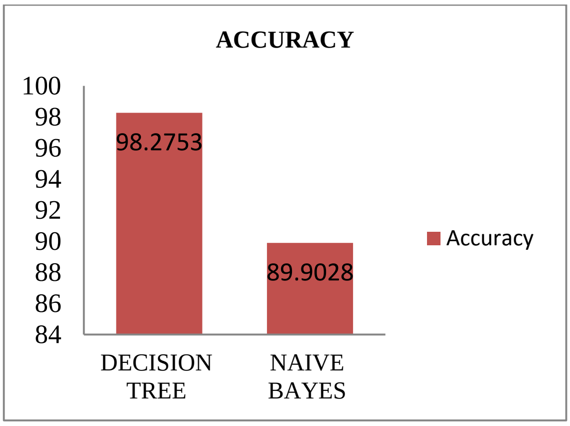
VI. PERFORMANCE EVALUATION

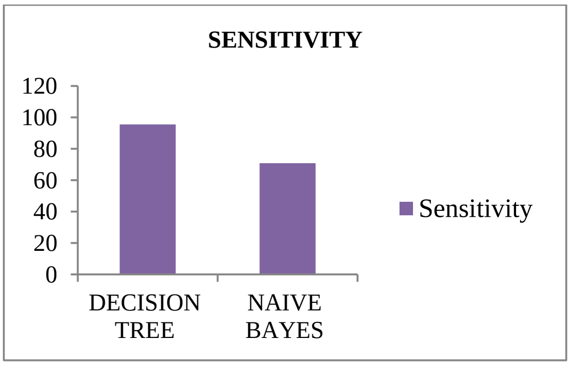
This is a measurement tool to calculate the performance:

- **Accuracy =** $\left[\frac{TP+TN}{TP+TN+FP+FN}\right]$
- **Sensitivity =** $\left[\frac{TP}{TP+FN}\right]$
- **Specificity =** $\left[\frac{TN}{TN+FP}\right]$

PERFORMANCE METRICS OF DT AND NB

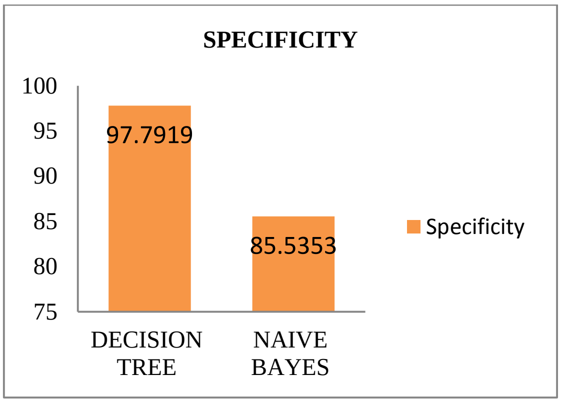
METHOD	DECISION TREE	NAIVE BAYES
Accuracy	98.2753	89.9028
Sensitivity	95.452	70.9042
Specificity	97.7919	85.5353





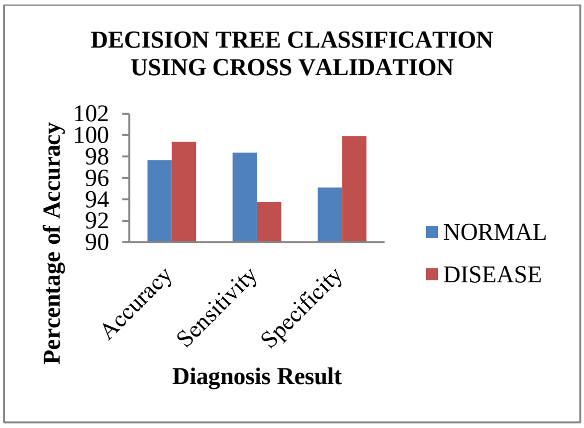
ISSN (ONLINE):2395-695X
ISSN (PRINT):2395-695X

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol.3, Issue.3, March 2017



SUMMARY OF THE CLASSIFICATION ACCURACY – DECISION TREE CLASSIFIER - CROSS VALIDATION

	NORMAL	DISEASE
Accuracy	97.6482	99.3885
Sensitivity	98.3686	93.7500
Specificity	95.1168	99.8974

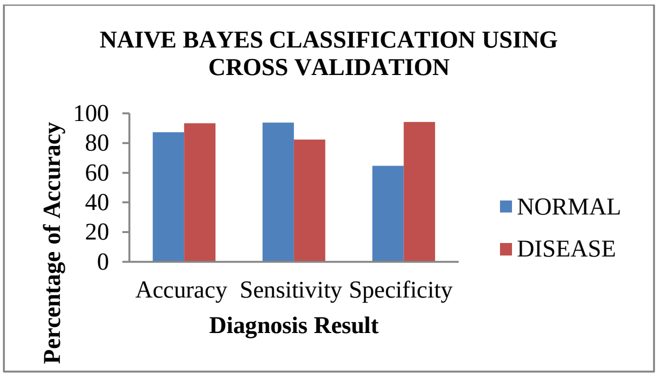


SUMMARY OF THE CLASSIFICATION ACCURACY – NAIVE BAYES CLASSIFIER - CROSS VALIDATION

	NORMAL	DISEASE
Accuracy	87.3001	93.3208
Sensitivity	93.7160	82.3864
Specificity	64.7558	94.3077

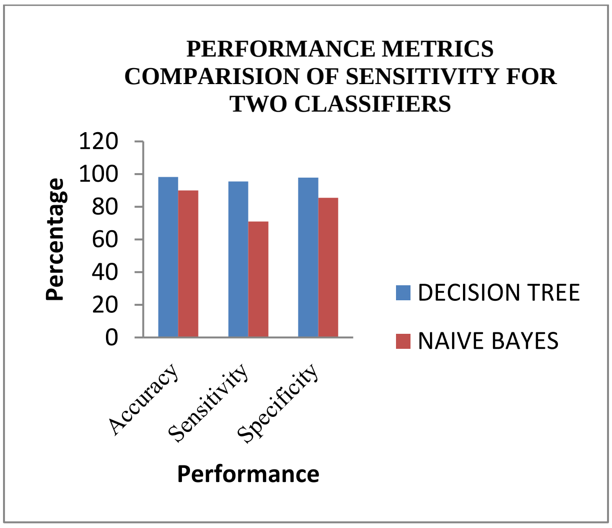
ISSN (ONLINE):2395-695X
ISSN (PRINT):2395-695X

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol.3, Issue.3, March 2017



PERFORMANCE ANALYSIS FOR FOUR CLASSIFIERS - 10 FOLD CROSS VALIDATION

METHOD	DECISION TREE	NAIVE BAYES
Accuracy	98.2753	89.9028
Sensitivity	95.4520	70.9042
Specificity	97.7919	85.5353



VII. CONCLUSION

The overall objective of our work is to predict more accurately the presence of heart disease. In this paper, UCI repository dataset are 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 provides accurate results as compare 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 predication e.g. Clustering, Time series, Association rules. The text mining can be used to mine huge amount of unstructured data available in healthcare industry database.

ISSN (ONLINE):2395-695X
ISSN (PRINT):2395-695X

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol.3, Issue.3, March 2017

REFERENCES

[1] Frawley and G. Piatetsky -Shapiro, Knowledge Discovery in Databases: An Overview. Published by the AAAI Press/ The MIT Press, Menlo Park, C.A 1996.

[2] Yanwei, X.; Wang, J.; Zhao, Z.; GAO, Y., "Combination data mining models with new medical data to predict outcome of coronary heart disease". Proceedings International Conference on Convergence Information Technology 2007, pp. 868 – 872.

[3] 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

[4] Niti Guru, Anil Dahiya, NavinRajpal, "Decision Support System for Heart Disease Diagnosis Using Neural Network", Delhi Business Review, Vol. 8, No. 1 (January - June 2007).

[5] HeonGyu Lee, Ki Yong Noh, KeunHoRyu, "Mining Bio signal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV," LNAI 4819: Emerging Technologies in Knowledge Discovery and Data Mining, pp. 56-66, May 2007.

[6] 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.

[7] Carlos Ordonez, "Improving Heart Disease Prediction Using Constrained Association Rules," Seminar Presentation at University of Tokyo, 2004.

[8] Kiyong Noh, HeonGyu Lee, Ho-Sun Shon, Bum Ju Lee, and KeunHoRyu, "Associative Classification Approach for Diagnosing Cardiovascular Disease", Springer, Vol: 345, pp: 721- 727, 2006.

[9] Franck Le Duff, CristianMunteanb, Marc Cuggiaa, Philippe Mabob, "Predicting Survival Causes After Out of Hospital Cardiac Arrest using Data Mining Method", Studies in health technology and informatics, Vol. 107, No. Pt 2, pp. 1256-9, 2004.

[10] LathaParthiban and R.Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", International Journal of Biological, Biomedical and Medical Sciences, Vol. 3, No. 3, 2008.

[11] Antepartum patient heart rate feature extraction and classification using empirical mode decomposition and support vector machine, Niranjana KrupaEmail author, Mohd Ali MA, Edmond Zahedi, Shuhaila Ahmed and Fauziah M Hassan, January 2011

[12] Performance Evaluation of K-Means and Heirarichal Clustering in Terms of Accuracy and Running Time, Nidhi Singh,.Divakar Singh, Department of computer Science & Engg.BUIT,BU,Bhopal.India.(M.P) , 2012.

[13] Determination of Patient State from Cardiotocogram Using LS-SVM with Particle Swarm Optimization and Binary Decision Tree, Ersen Yılmaz and Çağlar Kılıkçıer Electrical-Electronic Engineering Department, Uludag University, 16059 Gorukle, Bursa, Turkey, Received 26 June 2013; Accepted 6 September 2013.