IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 04, Issue 07 (July. 2014), ||V3|| PP 36-39

Comparative study of Data Mining Approaches for prediction Heart Diseases

Dr.Hari Ganesh S[1], Gajenthiran M[2]
[1]Asst. Professor, [2]M.Phil. Scholar,
Department of Computer Applications,
Bishop Heber College (Autonomous),
Trichirappalli-620 017

Abstract: - Data mining is the process of finding useful and relevant information from the databases. There are several types of data mining techniques are available. Association Rule, Classification, Neural Networks, Clusteringare some of the most important data mining techniques. Data mining process may take important role in Health care Industries. Most commonly the data mining process is used in health care industries for the process of prediction of diseases. This paper analysis the Heart Disease prediction approaches using classification technique. Here we are using three different kinds of classifiers named Naïve Bayes, Decision Table, and J48. The data set which is used here is taken from the UCI repository.Weka tool is used.

Keywords: - *Heart disease prediction, Classification, Decision Table, Naïve Bayes.*

I. INTRODUCTION

Data mining is the process of finding useful and relevant information from the various types of databases. The process of extraction of the hidden predictive information from the databases is also called as Data mining.[1] Nowadays data mining may take important role in several types of fields like Medical, Science, Railway etc. Association rule, Classification is one of the type of widely using data mining technique for prediction of heart diseases.

1.1 Classification

Classification is the process of categorization the data which are in the databases for its most effective and efficient use. For example, data can be broken down according to its topical content, file type, operating platform, average file size megabytes or gigabytes, when it was created, when it was last accessed or modified, which person or department last accessed or modified it, and which personnel or departments use it the most. A well-planned data classification system makes essential data easy to find. There are several types of classification models are available. But here we are using four different kinds of classification models. They are Naïve Bayes, Decision Tree, and J48.

1.2 Heart Diseases

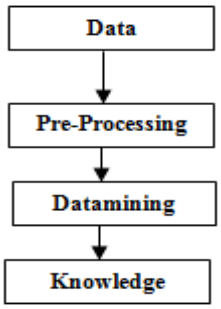
Heart disease is one of the type of disease which will affects the operations of the heart. Nowadays Heart disease is the major reason for deaths. There are several kinds of factors which increases the risk of Heart Diseases. Here we are using fourteen kinds of factors for the purpose predicting the heart disease. But the following are considered as important reasons for heart diseases.

- Age
- Blood Pressure
- Smoking habit
- Cholesterol
- Exercise and weight
- Blood Sugar
- Depression

The Heart Diseases are mostly caused by the above factors. The Survey result of World Health Organization says 12 million deaths are occurred worldwide, every year due to Heart diseases. If we want to find the solution for this problem, there are several kinds of prediction processes are needed. This paper will analyse various kinds of Heart Disease Prediction Techniques.

1.3 Weka Tool:

WEKA stands for Waikato Environment for Knowledge Learning. It was developed by the University of Waikato, New Zealand. WEKA supports many data mining tasks such as re-processing, classification, clustering, regression and future selection to name a few[2].The workflow of WEKA would be follows:



Naïve Bayes:

Naïve Bayes is a simple probabilistic classifier based on applying Bayes theorem with strong independence assumptions. A more descriptive term for underlying probability model would be "independent feature model". In simple terms, a Naïve Bayes classifier assumes that the presence of a particular feature of a class is unrelated to the presence of any other feature.[3] Depending on the precise nature of the probability model, Naïve Bayes classifiers can be trained very sufficiently in a supervised learning setting. In many practical applications, parameter estimation for Naïve Bayes models uses the method of maximum likelihood.

1.5. Decision Table:

The Decision tree is a structure that includes root node, branch and leaf node. Each internal node denotes a test on attribute, each branch denotes the outcome of test and each leaf node holds the class label [4]. The topmost node in the tree is the root node. There are three main advantages of decision tree.

- It does not require any domain knowledge.
- It is easy to assimilate by human.
- Learning and classification steps of decision tree are simple and fast.

1.6 J48:

The J48 Decision tree classifier follows the following simple algorithm. In order to classify a new item, it first needs to create a decision tree based on the attribute values of the available training data [5]. So, whenever it encounters a set of items (training set) it identifies the attribute that discriminates the various instances most clearly. This feature that ableto tell us most about the data instances so that we can classify them the best is said to have the highest information gain. Now, among the possible values of this feature, if there is any value for which there is no ambiguity, that is, for which the data instances falling within its category have the same value for that target variable, then we terminate that branch and assign to it the target value that we have obtained [6].

II. DATA DESCRIPTION

The data are collected from UCI repository. The objective of this data set is to analysis the Heart Disease based on the given attributes. The data set consists of 13 attributes that are used to predict the Heart Disease. The detail descriptions of the attributes are given as below.

Table 2.1: Attributes for Heart Disease

No	Name of the Attribute	Description
1	Age	Patient's age
2	Sex	Male/Female
3	Cp	Constrictive Pericarditis
4	Fbs	Fasting Blood Sugar
5	Trestbps	Resting Blood Sugar
6	Restecg	Resting Electrocardiograph Results
7	Thalach	Maximum Heart Attack achieved.
8	Exang	Exercise Induced Angina.
9	Oldpeak	ST depression by exercise relative to

		rest.
10	Slope	The slope of the peak exercise st segment.
11	Chol	Cholesterol
12	Ca	Number of major vessels (0-3) colored by fluoroscopy.
13	Num	Diagnosis of heart disease.

The attributes are given based on data types. The data set is based on the numeric and nominal data type.

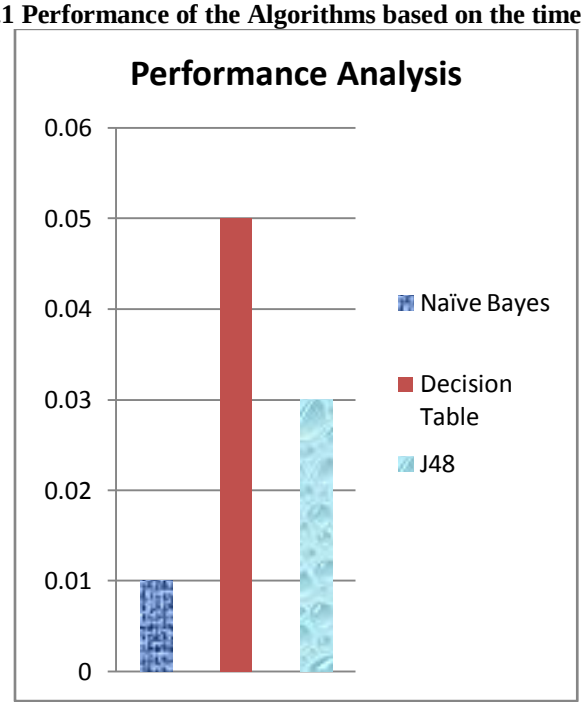
III. EXPERIMENTAL RESULTS

The given three types of algorithms like Naïve Bayes, Decision Table, and J48 are applied on the Heart Disease data set in WEKA and the performance of the algorithm are given based various factors. The performance can be obtained based on the time taken to build the tree and correctly classified instances.

Table 3.1 Time taken by the algorithms

Name of the Algorithm	Time Taken to build the decision tree
Naïve Bayes	0.01 seconds
Decision Table	0.05 seconds
J48	0.03 seconds

Fig 3.1 Performance of the Algorithms based on the time taken



X-Axis: Classification of Algorithms
Y-Axis: Time Range

The dataset consists of 303 instances and they are applied as a test case in the classification algorithms. The performance of the algorithms can be known from the instances that are correctly classified. The instances which are correctly classified using the WEKA tool can be given as below,

Table 3.2 Number of instances correctly classified

Name of the Algorithm	Number of correct instance	Accuracy
Naïve Bayes	253	83.4%
Decision Table	231	76.2%
J48	235	77.5%

IV. DISCUSSION

The above three algorithms predicts the class label. The final output will be patterns which are used to find out whether the person is affected by the Heart Disease or not. A Confusion Matrix is a useful visualization tool for analyzing the classifier accuracy. Structure of the confusion matrix can be given as below

Table 4.1 Structure of the Confusion Matrix

TP	**TN**
FP	**FN**

Where

- **TP** is True Positive: Heart Disease patients correctly identified as Heart Disease.
- **FP** is False Positive: Healthy people incorrectly identified as Heart Disease.
- **TN** is True Negative: Healthy people correctly identified as healthy.
- **FN** is False Negative: Heart Disease patients incorrectly identified as healthy.

The Confusion Matrix for the classification algorithms such as Naïve Bayes, Decision Table and J48 can be given as follows based on the execution of the algorithm using WEKA tool.

Table 4.2 Confusion Matrix for Naïve Bayes

143	**22**
28	**110**

Table 4.3 Confusion Matrix for Decision Table

133	**32**
40	**98**

Table 4.4 Confusion Matrix for J48

137	**28**
40	**98**

V. CONCLUSION

Data mining plays a major role in extracting the hidden information in the medical data base. The data pre-processing is used in order to improve the quality of the data. This model is built based as a test case on the UCI repository dataset. The experiment has been successfully performed with several data mining classification techniques and it is found that the Naïve Bayes algorithm gives a better performance over the supplied data set with the accuracy of 83.4%. It is believed that the data mining can significantly help in the Heart Disease research and ultimately improve the quality of health care of Heart Disease patients. It can also be implemented using several classification techniques.

REFERENCES

[1] Dharminder Kumar, Deepak Bhardwaj"Rise of Data Mining: Current and Future Application", *International Journal of Computer Science, Vol- 8, Issue 5, September-2011.*

[2] ZdravkoMarkovIngridRussell"An Introduction to the WEKA Data Mining System", *Zdravko Markov Central Connecticut State University.*

[3] VikasChaurasia, SaurabhPal"Early Prediction of Heart disease using data mining techniques",*Carib.j.SciTech,2013,Vol.1,208-217*

[4] NidhiBhatla, KiranJyoti"An Analysis of Heart Disease Prediction using Different Data Mining Techniques",*International Journal of Engineering Research & Technology (IJERT).*

[5] Jay Gholap"Performance Tuning Of J48",Dept. of Computer Engineering College of Engineering, Pune, Maharashtra, India.

[6] Shadab Adam Pattekari and AsmaParveen "Prediction System for Heart Disease Using Naive Bayes", *International Journal of Advanced Computer and Mathematical Sciences. Vol 3, Issue 3, 2012, pp 290-294.*