iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 4 **Issue:** II **Month of publication:** February 2016

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Heart Disease Prediction Using Data Mining Classification

K.Gomathi[1], Dr. Shanmugapriyaa[2]
[1]*Research Scholar, Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore*
[2]*Professor, Department of Information Technology, Karpagam Academy of Higher Education, Coimbatore*

Abstract: In this paper we present an analysis of the Heart disease for male patients using data mining techniques. The preprocessed data set consists of 210 records, which have all the available 8 fields from the database. We have investigated three data mining techniques: the Naïve Bayes, Artificial neural network, and the J48 decision tree algorithms. Our Analysis Shows that of these three classification models Naïve Bayes predicts heart disease with higher Accuarcy.
Keywords: Data Mining, Heart disease, Naive Bayes, ANN, Decision tree.

I. INTRODUCTION

Data mining is process of extracting useful information from large amount of databases. Data mining is most useful in an exploratory analysis because of nontrivial information in large volumes of data. Data mining is the process of extracting data for finding buried patterns which can be transformed into significant. Data mining knowledge afford a user-oriented approach to new and concealed patterns in the data. The knowledge which is exposed can be used by the healthcare practitioners to get better quality of service and to reduce the extent of adverse medicine effect. Hospitals have to reduce the charge of medical tests. They can attain these consequences by employing suitable decision support systems. Health care data is enormous. It consists of patient centric data, resource organization data and altered data. Medical care organizations must have capability to explore data. Treatment records of millions of patients can be hoarded and data mining techniques will aid in answering numerous essential and decisive questions interrelated to health care. Data mining techniques has been performed in healthcare domain. This realization is in the arouse of explosion of difficult medical data. Medicinal data mining can utilize the veiled patterns present in huge medical data which otherwise is left undiscovered. Data mining techniques which are useful to medical data include association rule mining for finding frequent patterns, prediction, classification and clustering. Data mining techniques are more useful in predicting heart diseases, breast cancer lung cancer, diabetes and etc.

II. HEART DISEASES

Heart disease is the leading cause of death in the U.S. At some point in your life, either you or one of your loved ones will be forced to make decisions about some aspect of heart disease. Knowing something about the anatomy and functioning of the heart, in particular how angina and heart attacks work, will enable you to make informed decisions about your health. Heart disease can strike suddenly and require you to make decisions quickly.

A. Heart Disease Facts

1) Heart disease is the leading cause of death for both men and women. More than half of the deaths due to heart disease in 2009 were in men.
2) About 610,000 Americans die from heart disease each year—that's 1 in every 4 deaths.
3) Coronary heart disease is the most common type of heart disease, killing more than 370,000 people annually.[1]
4) In the United States, someone has a heart attack every 43 seconds. Each minute, someone in the United States dies from a heart disease-related event.[2]
5) Heart disease is the leading cause of death for people of most racial/ethnic groups in the United States, including African Americans, Hispanics, and whites. For Asian Americans or Pacific Islanders and American Indians or Alaska Natives, heart disease is second only to cancer.
6) Coronary heart disease alone costs the United States $108.9 billion each year. This total includes the cost of health care services, medications, and lost productivity.

B. Risk Factors

High blood pressure, high LDL cholesterol, and smoking are key heart disease risk factors for heart disease. About **half of Americans** (49%) have at least one of these three risk factors.

Several other medical conditions and lifestyle choices can also put people at a higher risk for heart disease, including:

1) Diabetes
2) Overweight and obesity
3) Poor diet
4) Physical inactivity
5) Excessive alcohol use

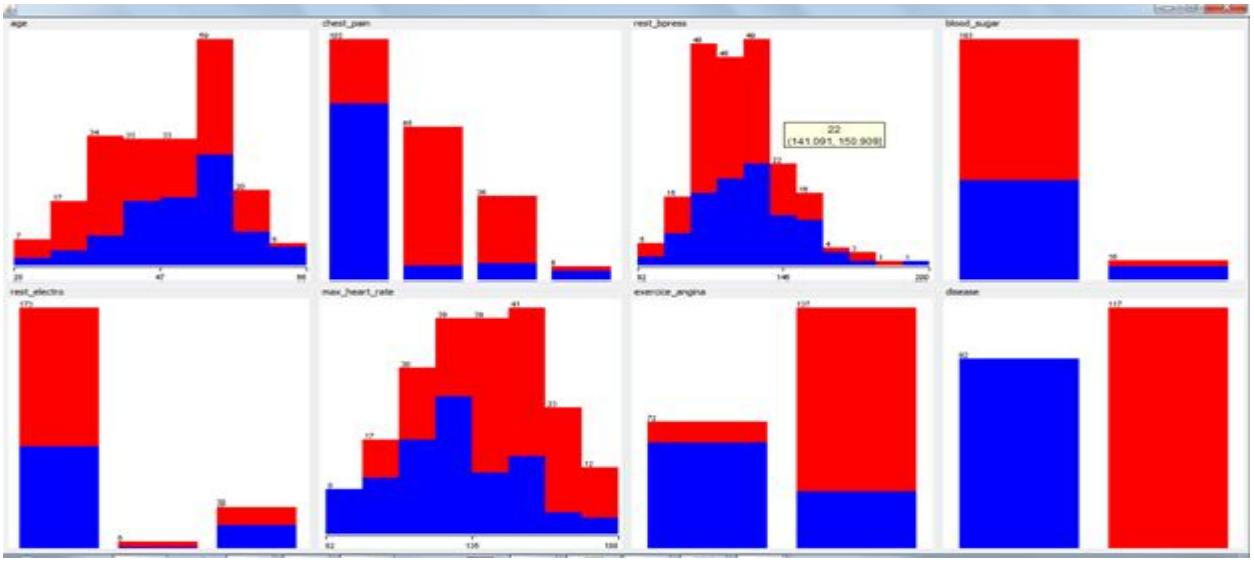
This paper analyzes the heart disease predictions using classification algorithms. These hidden patterns can be used for health diagnosis in medicinal data. Data mining technology afford an effective approach to latest and indefinite patterns in the data. The information which is identified can be used by the healthcare administrators to get better services. Data mining classification techniques like J48, Decision Trees, and Naive Bayes are used to analyze the dataset based on disease attribute.

III. METHODOLOGY

In order to carry out experimentations and implementations weka was used as the data mining tool. Weka (Waikato Environment for Knowledge Analysis) is a data mining tool written in java developed at Waikato. WEKA is a very good data mining tool for the users to classify the accuracy on the basis of datasets by applying different algorithmic approaches and compared in the field of bioinformatics. Explorer, Experimenter and Knowledge flow are the interface available in WEKA that has been used by us. In this paper we have used these data mining techniques to predict the survivability of dengue disease through classification of different algorithms accuracy.

It has four applications:

A. Explorer: The explorer interface has several panels like preprocess, classify, cluster, associate, select attribute and visualize. But in this interface our main focus is on the Classification Panel
B. Experimenter: This interface provides facility for systematic comparison of different algorithms on basis of given datasets. Each algorithm runs 10 times and then the accuracy reported.
C. Knowledge Flow: It is an alternative to the explorer interface. The only difference between this and others is that here user selects Weka component from toolbar and connects them to make a layout for running the algorithms.
D. Simple CLI: Simple CLI means command line interface. User performs operations through a command line interface by giving instructions to the operating system. This interface is less popular as compared to other three.

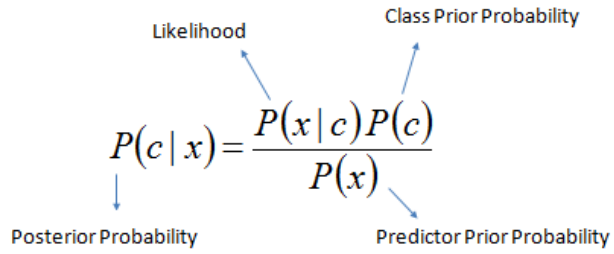


IV. DATA MINING TECHNIQUES

A. Naive Bayes

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.

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.



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

1) $P(c|x)$ is the posterior probability of *class* (*target*) given *predictor* (*attribute*).
2) $P(c)$ is the prior probability of *class*.
3) $P(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
4) $P(x)$ is the prior probability of *predictor*.

B. Artificial Neural Network

A multilayer perceptron is a feedforward artificial neural network model that maps sets of input data onto a set of appropriate output. It is a modification of the standard linear perceptron in that it uses three or more layers of neurons (nodes) with nonlinear activation functions, and is more powerful than the perceptron in that it can distinguish data that is not linearly separable, or separable by a hyper-plane.

C. Decision Tree (J48)

Decision tree models are commonly used in data mining to examine data and induce the tree and its rules that will be used to make predictions[10].The prediction could be to predict categorical values (classification trees) when instances are to be placed in categories or classes. Decision tree is a classifier in the form of a tree structure where each node is either a leaf node, indicating the value of the target attribute or class of the examples, or a decision node, specifying some test to be carried out on a single attribute-value, with one branch and sub-tree for each possible outcome of the test. A decision tree can be used to classify an example by starting at the root of the tree and moving through it until a leaf node is reached, which provides the classification of the instance.

V. EXPERIMENTAL RESULTS

In this study, the accuracy of three data mining techniques is compared. The goal is to have high accuracy, besides high precision and recall metrics. Although these metrics are used more often in the field of information retrieval, here we have considered them as they are related to the other existing metrics such as specificity and sensitivity. These metrics can be derived from the confusion matrix and can be easily converted to true-positive (TP) and false-positive (FP) metrics.

A. Confusion Matrix for three classification methods

1) Confusion Matrix for Naïve Bayes

	a	b
a	67	25
b	17	100

2) Confusion Matrix for ANN

	a	b
a	64	28
b	21	96

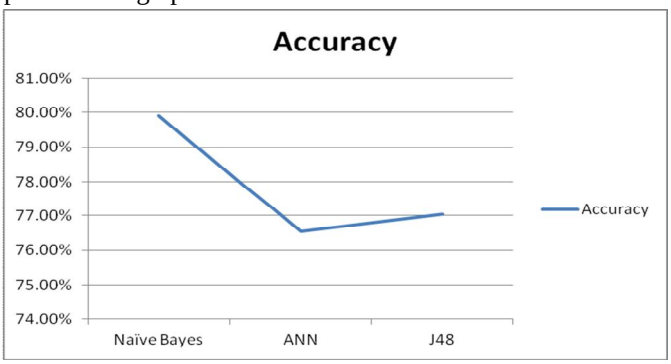
3) Confusion Matrix for J48

	a	b
a	66	26
b	22	95

Table Shows accuracy for three different classifications

Classification	Accuracy	Time Taken
Naïve Bayes	79.9043%	0.01 Seconds
ANN	76.555 %	1.55 Seconds
J48	77.0335%	0.01 Seconds

The Accuracy is of each method is plotted on a graph as below:

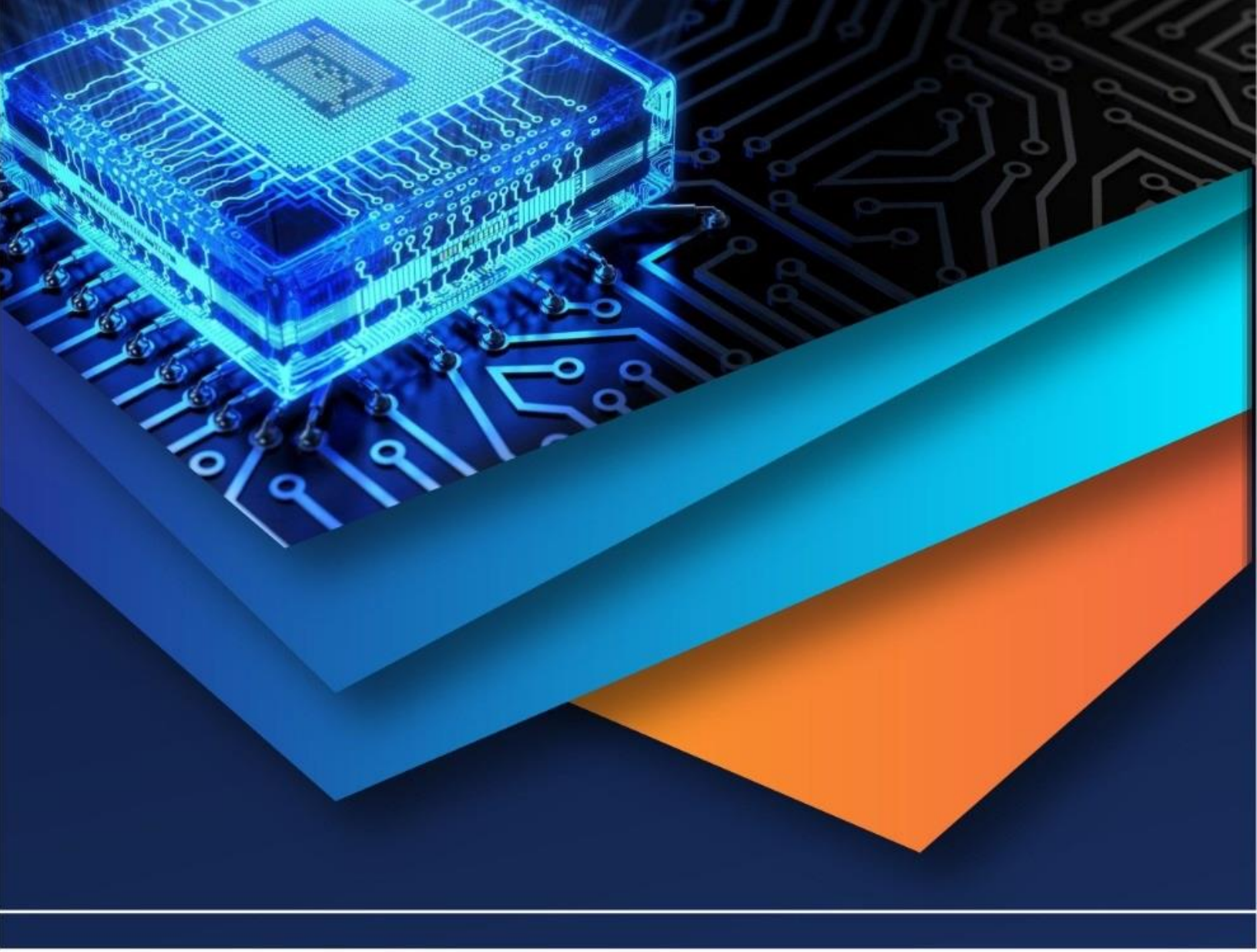


VI. CONCLUSIONS

In this paper we have discuss some of effective techniques that can be used for heart diseases classification and the accuracy of classification techniques is evaluated based on the selected classifier algorithm. An important challenge in data mining and machine learning areas is to build precise and computationally efficient classifiers for Medical applications. The performance of Naive Bayes shows high level compare with other classifiers.

REFERENCES

[1] http://www.cdc.gov/dhdsp/data_statistics/fact_sheets/fs_heart_disease.htm
[2] Chaitrali S.Danagre, Sulabha S.Apte, Ph.D, Improved Studyof Heart Disease Prediction Systemusing Data mining Classification Techniques,IJCA,June 2012.
[3] Sellappan Palaniappan,Rafiah Awang,"Intelligent Heart Disease Prediction System using data MiningTechniques",IJCSNS,Vol.8 No.8,August 2008
[4] Heartdisease Male:http://archieve.ics.uci.edu/ml/datasets/Heart+disease+male
[5] Witten H.I., Frank E., Data Mining: Practical Machine Learning Tools and Techniques, Second edition, Morgan Kaufmann Publishers, 2005.s
[6] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques, 2nd Edition. San Fransisco:Morgan Kaufmann; 2005



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY