International Journal of Management, IT & Engineering
Vol. 9 Issue 7, July 2019,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

HEART DISEASE PREDICTION USING CLASSIFICATION TECHNIQUES: A COMPARATIVE STUDY

Nikhil Ghodekar*

Prof. Priya Chandran*

Prof. ShravaniPawar*

Abstract

As per WHO around 20 million people die around the world due to cardiovascular disease. If proper medical care is provided, then many lives can be saved. The cause of heart attack in the health sciences sector is a difficult task. Large number of health care data is available for analysis. Different data mining algorithms are successfully applied to predict heart disease.Some of the data mining and machine learning techniques are used to predict the heart disease, such as Multilayer perceptron Neural Network (MLPNN), Decision tree, Naïve Bayes. This research paper compares three classification approaches, naive bayes, J48 and MLPNN, to predict heart disease from the give data set and studies the effectiveness of each method. Out if the three methods studied, MLPNN resulted with maximum correctly classified instances of data with a minimum root mean square error value, 0.14.

Keywords:

heart disease prediction;

MLPNN;

J48;

naive bayes;

classification;

data mining.

* **Student, MCA Dept.BVIMIT, Navi Mumbai.**

1. Introduction

According to a recent survey by WHO organisation 20 million people die each year. A heart is a major organ that compares to the brain, which is a priority in the human body. It pumps the blood and supplies to all organs of the whole body. In 2015, a total of 17.7 million deaths worldwide have been caused by human heart disease. An estimated 7.4 million people's deaths out of total deaths were due to coronary heart sickness and estimated 6.9 million deaths have been due to heart stroke [10]. It will increase to 75 million in the year 2030 [7][8]. Medical professionals working in the field of heart disease have their own limitation; they can predict chance of heart attack up to 67% accuracy [9].

We know that heart disease is top most reason for deaths of humans all across the globe. Every one of the five deaths are due to heart disease. Almost 1/2 folks residents of US have minimum one risk component for heart sickness along with high blood strain, obesity, sedentary lifestyle, high blood sugar etc.

A cardiovascular disease usually refers to conditions in which blood vessels are narrowed or blocked, which can cause heart attack, chest pain (angina) or stroke. A heart arrhythmia is an abnormal heartbeat.

Heart arrhythmia symptoms can be [9]:

- Fluttering in your chest
- Racing heartbeat (tachycardia)
- Slow heartbeat (bradycardia)
- Chest pain or discomfort
- Shortness of breath
- Light-headedness

The use of information technology in the health service industry is increasing day by day to help doctors in decision-making activities. Medical organizations around the world collect information on various health problems. This data can be exploited using various data learning techniques to get useful information. But the collected data is very large and often this data can

be very messy. Thus, these algorithms have become very useful, in recent times, to predict the presence or absence of heart related diseases accurately.

Data mining is a knowledge discovery technique. It analyzes data and encapsulates it into useful information. It is the process of sorting through large data sets to identify patterns and establish relationships to solve problems through data analysis. A Classification parameter looks for new patterns. Classification algorithms tend to predict variables based on other factors within the database.

2. Literature review

Several data mining techniques are being used by researchers of different early disease predictions. Large volume of data is available in health care for this study. Many studies have been done to predict heart disease in an early stage. Researchers have implemented various data mining techniques to diagnose heart disease.

Sellappan P et al. [5] proposed an Intelligent Heart Disease Prediction System (IHDPS) and is developed using data mining techniques Naive Bayes, Neural Network, and Decision tree. Each method has its own power to get appropriate results. To build this system they used hidden patterns and relationship between them. It is web-based, user-friendly and expandable.

To develop the multi-parametric feature with linear and nonlinear characteristics of HRV (Heart Rate Variability), a novel technique was proposed by HeonGyu Lee et al. [6]. They have used several classifiers to analyse the data. Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques proposed by Chaitrali S. Dangare and SulabhaS.Apte. [1]. The authors used navies Bayes, decision tree, Neural network algorithm to predict heart disease and also analysed the accuracy of algorithms. A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach is done by M. Marimuthu et. al [2]. Using ANN, Fussy Logic, KNN and SVM. Their approach used various attributes such as age, sex fasting blood pressure, chest pain type, chest pain type, resting ecg, number of vessels, serum cholesterol, thalach, ST depression, painloc, fasting blood sugar, smoke, hypertension, food habits, weight, height and obesity. The authors have done analysis using weka software.

Tanvi Sharma et.al [3] proposed Intelligent Heart Disease Prediction System using Machine Learning. They used classification algorithm of machine learning techniques for diagnosis of cardio vascular and heart diseases.

V.V. Ramalingam et. al. [4] presents a survey of various models based on data mining algorithms and techniques and analyzed their performance. Models based on supervised learning algorithms such as Support Vector Machines (SVM), K-Nearest Neighbour (KNN), Naïve Bayes, Decision Trees (DT), Random Forest (RF) and ensemble models are found very popular among the researchers.

In our research we have studied heart disease prediction using three algorithms, Multilayer perceptron, J48 and Naive Bayes. We also analysed the effectiveness and accuracy of each algorithm using the same data set.

3. Research Method

The main objective of the classification issue is to identify the category / classes.The datamining work in this research is to create models for predictors of class based on selected properties. The research applies the following algorithms: J48, Naive Bayes and Multilayer perceptron algorithm to classify and develop a model to diagnose heart disease.

3.1Multi-Layered Perceptron

The neural network is an Artificial Neural Network (ANN), often called "neural network" (NN), a mathematical model or computational model that is based on the biological nervous network. The term Multi-layered Perceptron is used with the input layer, one or more hidden layers and the structure of an output layer for the neural network. Each of the layers has an interconnected assembly of ordinary Processing elements called neurons. These processing elements are organized in a layered fashion. In each layer, each neuron is attached to the posterior layer and so on. The correlation between layers is called weight. Despite their simplified structure, through the learning and generalization of neuralnetworks, there is theability to mimic human characteristics to solve the problem. MLP can be used to model non-linear systems due to their ability to learn system behavior under inspection by samples. For the successful application of

the MLP neural network, to meet the required performance criteria, a person should determine the internal parameters such as initial load and network structure. The primary function of neurons of the input layer is to divide input xi into neurons in the hidden layer. The neuron of the hidden layer adds input signal xi with weights wji of respective connections from the input layer. An MLPNN with one hidden layer is shown in figure 3.1.1.

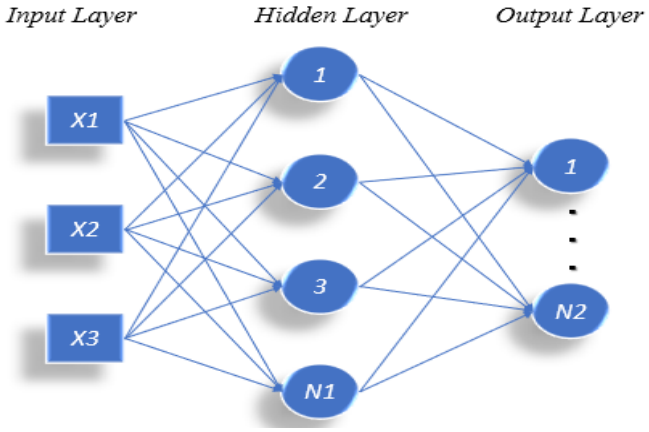

Figure 3.1.1: Multilayer perceptron

3.2 J48 algorithm

The J48 algorithm is called the optimized implementation of C4.5 or better version of C4.5. An Output is the decision tree given by J48. J48 is an extension of ID3. The diagnosis tree is the same as if there are different nodes in the tree structure, such as root nodes, intermediate nodes, and leaf nodes. There is a decision in every node in the tree and this decision leads to our results because the name is the decision tree. Decision tree divides the input space of the data set in the interconnected areas, where there is an action to describe or expand a label, a value or its data points in each field. The criterion for splitting is used to calculate in the decision tree that what is thebest feature to split the tree of that part of the training data reaching a particular node.

Some basic steps are given below to construct tree: -

I. First, check whether all cases belongs to same class, then the tree is a leaf and is labeled with that class.

II. For each attribute, calculate the information and information gain.

III. Find the best splitting attribute (depending upon current selection criterion).

Counting information gain:

"Entropy" is used in this process. Entropy is a measure of disorder of data. Entropy is measured in bits, nats or bans. Thisis also called measurement of uncertainty in any random variable

The information gain for a particular attribute say A at a node is calculated as under:

$$Information\ Gain(N, A) = Entropy(N) - \sum_{values(A)} \frac{|N_i|}{|N|} Entropy(N)$$

Where N is set of instances at that particular node and |N| is its cardinality, Niis the subset of N for which attribute A has value i, and entropy of the set N is calculated as:

$$entropy(N) = \sum_{i=1}^{No.of\ Classes} -P_i log2P_i$$

Where Piis proportion of instances in N that have their ith class value as output attributes.

3.3 Naive Bayes

Naive Bayes use the probabilistic Naïve Bayes classifier. Naive Bayes Simple uses the normal distribution to model numeric attributes. They can predict class membership probabilities, such as the prob that a given tuple belongs to a particular class.Here we have a hypothesis that the given data belongs to a particular class.Then calculate the probability of hypothesis to be true. Naïve Bayes is an incremental version that processes one request at a time. It can use a kernel estimator but not discretization.

$P(c|x)=\underline{P(x|c)P(c)}$

$P(x)$

$(c|X)=P(x1|c)\times P(x2|c)\times\ldots\times P(xn|c)\times P(c)$

Where

$(c|x)$is the posterior probability of class given predictor

$(c)$is the prior probability of class

$(x|c)$is the likelihood probability of predictor given class

$(X)$is the prior probability of predictor

4. Results and Analysis

We have carried out the research in weka software and analysed the performance of three data mining techniques. We have used heart disease data set [11] to estimate the potential cardiology of patient datasets and to determine which model determines the highest percentage of correct estimates. The dataset consists of total 100 records in heart disease database. The dataset consists of total 100 records in Heart disease database. A confusion matrix is obtained to calculate the accuracy of classification. Confusion matrix shows how many instances have been given to each class. The details of the parameters considered for the research study are given below:

Heart disease dataset

This database contains 13 attributes:

Attribute Information:

1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholesterol in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina
10. old peak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) coloured by fluoroscopy
13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

Dataset analysis

It is important to review our data set before modelling starts. We have analysed the distribution of each attribute and their interactions between attributes. We have noticed few points:

- There are 506 instances. If we use 10-fold cross validation later to evaluate the algorithms, then each fold will be comprised of about 50 instances, which is fine.
- There are 14 attributes, 13 inputs and 1 output variable.

The attributes and their distributions and interactions are shown in figure 4.1, figure 4.2 and figure 4.3 respectively.

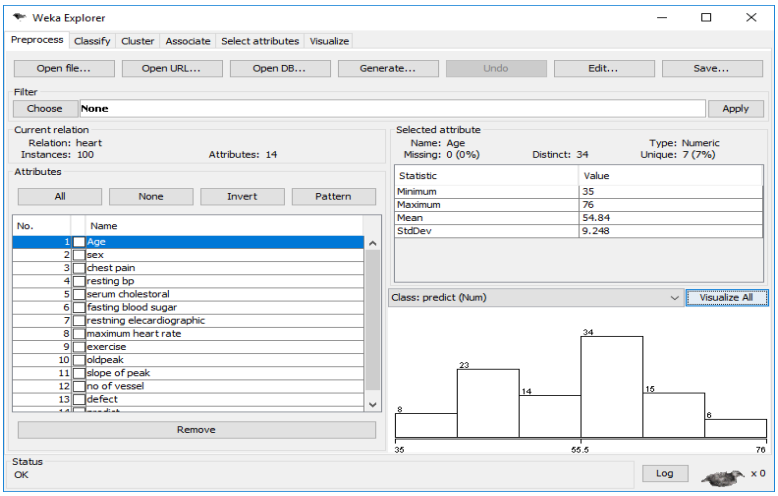

Figure 4.1: Dataset

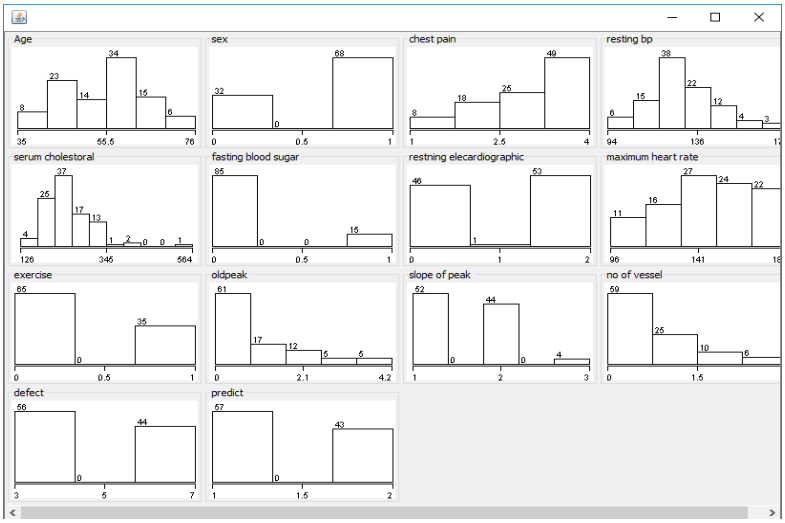

Figure 4.2: Attribute Distribution

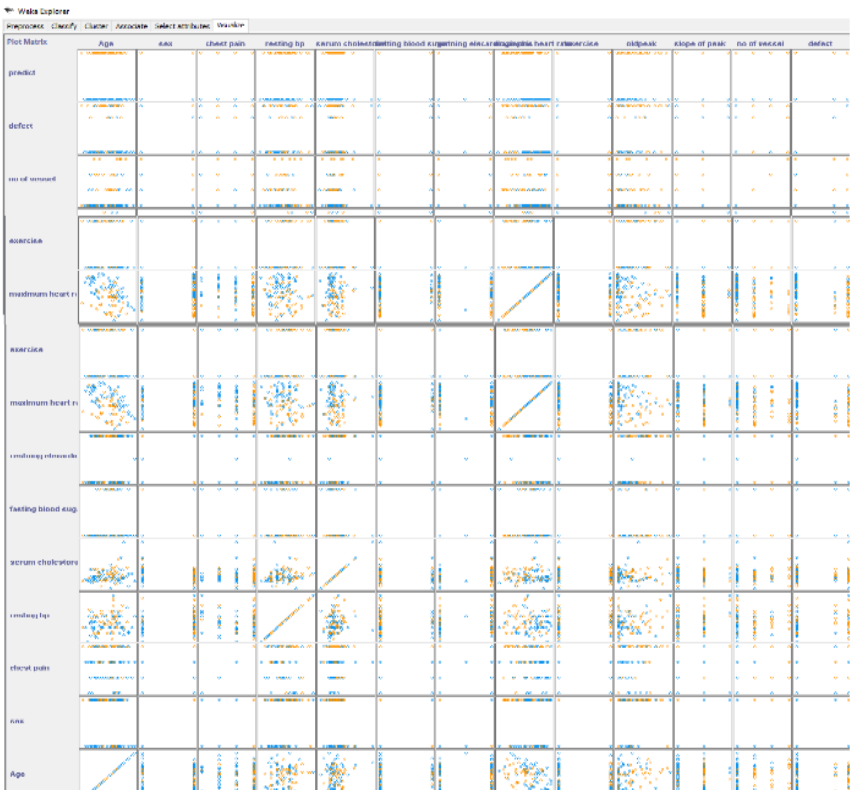

Figure 4.3: Attribute Interaction

In our experiment we have two classes, and therefore we have a 2x2 confusion matrix, which is shown in table 4.1.

Class a = YES (has heart disease)

Class b = NO (no heart disease)

Table 4.1: 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.

Results obtained with 13 attributes are specified in table 4.2a,4.2b and 4.2c respectively

Table 4.2a: Confusion Matrix of Naive Bayes Algorithm

	A	B
A	57	9
B	8	26

Table 3.2b: Confusion Matrix of J48Algorithm

	a	B
A	64	2
B	5	29

Table 4.2c: Confusion Matrix of MLP Algorithm

	a	B
A	66	0
B	2	32

Table 4.2: Predictive Performance of classifier

Evaluation Criteria	Classifier		
	Naive Bayes	J48	MLP
Timing to Build Model (sec)	0	0.01	0.36

Correctly classified Instance	83	93	98
Incorrectly Classified Instance	17	7	2
Predictive Accuracy	83	93	98
Kappa statistics	0.62	0.84	0.95
Mean Absolute error	0.28	0.12	0.03
Root mean square	0.35	0.25	0.14
Relative absolute error	61.17	26.93	7.54
Root relative squared error	74.65	51.95	30.24

The results of the comparative study are listed in table 4.2. The effectiveness of MLPNN, naive bayes and J48 are measured using the listed parameters. Out of the three methods which we have used for prediction, 83, 93 and 98 are the correctly classified instances of naïve Bayes, J48 and MLP methods respectively. The Root mean square error is 0.35, 0.25 and 0.14 respectively.

The following figures show the performance of three algorithms. Correctly classified instances and Root mean square of each method is shown in figure4.4 and figure 4.5 respectively.

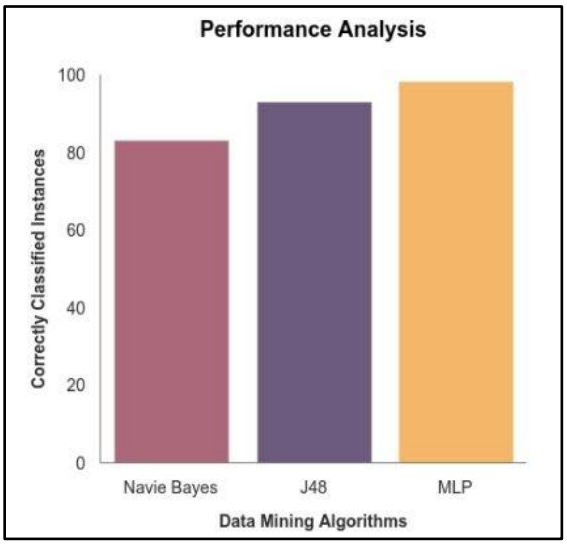

Figure 4.4: Correctly classified instances

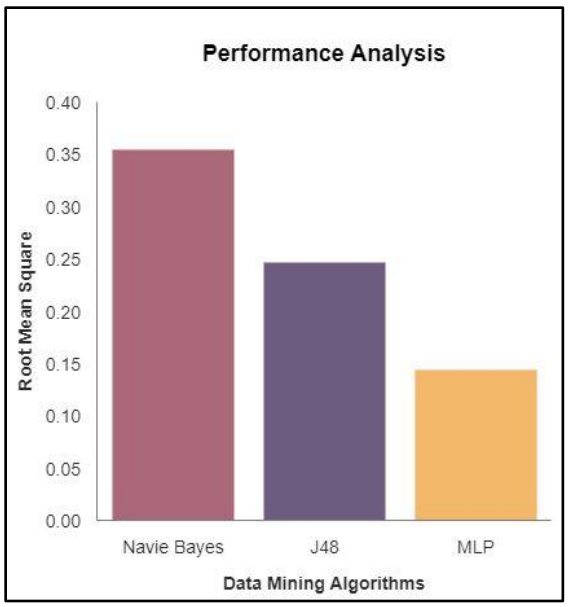

Figure 4.5: Root Mean Square

5. Conclusion

In our research study, we have implemented naive bayes, J48 and MLP neural network data mining classification techniques. Separate experiments were conducted in each type: the first one is to measure the performance of the decision tree classifier; the second one is to measure the performance of the naïve Bayes classifier, the third one to measure the performance of the neural network. From results it has been seen that MLP neural network provides accurate results as compare to naive bayes and J48 algorithm.

References

[1] Chaitrali S. Dangare, Sulabha S. Apte "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques" *in 2012 International Journal of Computer Applications (0975 – 888).*

[2] M. Marimuthu, M. Abinaya, K. S. Hariesh, K. Madhankumar "Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach" *in 2018 International Journal of Computer Applications (0975 – 8887) Volume 181 – No. 18, September 2018.*

[3] Tanvi Sharma, Sahil Verma, Kavita "Intelligent Heart Disease Prediction System using Machine Learning: A Review**"** *in 2017International Journal of Recent Research Aspects ISSN: 2349-7688, Vol. 4, Issue 2, June 2017, pp. 94-97.*

[4] V.V. Ramalingam, AyantanDandapath, M Karthik Raja "Heart disease prediction using machine learning techniques: a survey" *International Journal of Engineering & Technology, 7(2.8) 2018 684-687.*

[5] SellappanPalaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", *IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.8, August 2008.*

[6] HeonGyu Lee, Ki Yong Noh, Keun Ho Ryu, "Mining Biosignal 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.*

[7] Himanshu Sharma,M A Rizvi" Prediction of Heart Disease using Machine Learning Algorithms: A Survey"*International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169,Volume: 5 Issue: 8.*

[8] William Carroll; G. Edward Miller, "Disease among Elderly Americans: Estimates for the US civilian non institutionalized population, 2010," *Med.Expend. Panel Surv., no. June, pp. 1–8, 2013.*

[9] V. Kirubha and S. M. Priya, "Survey on DataMining Algorithms in DiseasePrediction," *vol. 38, no. 3, pp. 124–128, 2016.*

[10] Kadam, Kalyani, Pooja Vinayak Kamat, and Amita P. Malav. "Cardiovascular Disease Prediction Using Data Mining Techniques: A Proposed Framework Using Big Data Approach." *Coronary and Cardiothoracic Critical Care: Breakthroughs in Research and Practice. IGI Global, 2019. 246-264.*

[11] http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/heart/