Prediction of Heart Disease using Supervised Learning Algorithms

T. Marikani
Assistant Professor,
Department of Computer Science,
Sree Muthukumaraswamy College,
Chennai – 118, India

K. Shyamala
Associate Professor,
PG & Research Department of Computer Science,
Dr.Ambedkar Govt. Arts and Science College,
Vyasarpadi, India

ABSTRACT
The diagnosis of disease is difficult but critical task in medicine. Data mining is the process of extracting hidden interesting patterns from massive database. In the healthcare industry it plays a significant task for predicting the disease. Heart disease is a single largest cause of death in developed countries and one of the main contributors to disease burden in developing countries. Data mining is a more convenient tool to assist physicians in detecting the diseases by obtaining knowledge and information regarding the disease from patient's data. By using data mining techniques it takes less time for the prediction of the disease with more accuracy. This paper aims at analyzing the various data mining techniques namely Decision Trees, Naive Bayes, Neural Networks, Random Forest Classification and Support Vector Machine by using the Cleveland dataset for Heart disease prediction. Few of the supervised learning algorithms are used for the prediction of heart disease. It provides a quick and easy understanding of various prediction models in data mining and helps to find the best model for further work.

Keywords
Data mining, Supervised learning, Decision trees, Naïve Bayes, Neural Network, SVM.

1. INTRODUCTION
Heart disease is the major causes of death among the people now-a-days. According to World Health Organization (WHO) estimation nearly 12 million of death occurs due to heart disease [4]. The WHO estimates almost the death rate would be increases to 23.6 million by 2030. In order to reduce the risk of heart disease, prediction should be done. Some of the attributes which predict the heart disease are age, sex, chest pain type, family history, ECG reading, cholesterol, blood sugar level [16].

Data mining helps to extract useful information and generate relationship among the attributes. Data mining is a discipline to realize knowledge from database. It is essential to find the best fit classification algorithm that has great accuracy on classification in the case of heart disease.

2. LITERATURE SURVEY
Heart disease is one of a major health problem in today's life. In this paper, various data mining techniques introduced in recent years for heart disease prediction are analyzed. Many authors investigated on this area and applied various data mining techniques. In this section, some of their work along with the accuracy is shown in Table 1.

Table 1: Various Supervised Learning Algorithm and its Accuracy

Sl. No	Author	Year of Publication	Techniques/ Methods used	Accuracy	Tools Used
1.	Resul Das et.al.,	2008	Neural Network in classification	89.01%	SAS enterprise miner 5.2
2.	M.Anbarasi et.al.,	2010	Naïve Bayes Decision tree Classification via clustering Naïve Bayes	96.5% 99.2% 88.3%	Weka
3.	A.Q.Ansari et,al.,	2011	Neuro-fuzzy intergrated system	Maximum accuracy has been obtained	Weka
4.	Chaitrail et.al.,	2012	Neural Network	100%	Weka
5.	T.John et.al.,	2012	Navie Bayes Multilayered Feed forward	85.18% 78.88%	Weka
6.	Nidhi Bhatla et.al.,	2012	Decision tree Naïve Bayes ANN	89% 86.53% 85.53%	Weka

7.	Jesmin Nahar et.al.,	2013	Computational Intelligence	Expect result can be produced by implementing data in naïve bayes.	Weka
8.	K.Thenmozhi et.al.,	2014	Naïve Bayes Decision tree Neural Network	94.44% 96.60% 99.25%	Weka
9.	Deepali Chandna	2014	Adaptive Neuro-fuzzy inference system (ANFIS)	98.24%	Weka and MATLAB
10.	Tamilarasi et.al.,	2015	Naïve Bayes KNN J48 CART ANN	85.92% 100% 91.85% 95.92% 99.25%	Weka
11.	K.Aravinthan et.al.,	2016	Naïve Bayes J48 ANN	81.3% 80.09% 82.56%	Weka

3. METHODOLOGY

The task of data mining in this research is to find out the best algorithm for the prediction of heart disease. The research applied in the following supervised learning algorithms like Decision tree, Naïve Bayes, Random Forest Tree, KNN and Support Vector Machine. These algorithms are used to classify and develop a model to diagnosis heart disease in the patients.

The objective of this research is to predict heart disease from the patient dataset using data mining techniques and to determine which model gives the better percentage of accuracy in the prediction of disease.

3.1. Data Source

The data set used for this work is from UCI Machine Learning repository in which the Cleveland heart disease dataset is used. The dataset has 303 instance and 76 attributes. However only 14 attributes are used in this paper. These 14 attributes are the mostly consider factors for the heart disease prediction [16]. Even though it has 303 instances of which only 297 are complete and the remaining rows contain missing values and removed from the experiment.

3.2. Proposed Framework

The classification algorithm in data mining is used to analysis the heart disease prediction. In this research work we used to deal with the data mining problem which is concerned with prediction. Data Mining is one of the most common tasks used to build models for the prediction of the class of an object on the basis of its attributes. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithms. In classification, test data are used to estimate the accuracy of the classification rules.

4. RESULTS AND DISCUSSION

In this section the collected heart disease dataset is implemented in the various classification algorithms like decision tree, Naïve Bayes, Neural Network and Support Vector Machine. Orange is the tool used for the experiment. It is a component-based visual programming for data mining, machine learning and data analysis.

The parameters/attributes used in the research work are: A1 – Age, A2 – Gender, A3 – CP (Chest pain), A4 – Resting blood pressure, A5 – cholesterol, A6 – fast blood sugar > 120 (1,0) A7 – rest ECG(normal, abnormal), A8 – Max HR(Heart Rate), A9 –Exercise induced angina (0,1), A10 – ST by exercise(0.5,2.5,1.6,2.5….) A11 – slope peak exercise ST, A12 – Major vessel colored(3.00,2,0,1,3…), A13 – thal(normal, fixed defect, reversible defect-heart status), A14 – Diameter Narrowing (0,1).

The following figure shows, how the classification algorithms are created in orange tool. This tool contains various widgets. Among those widgets, we select the classify widget to perform the experiment and it is shown in the following Fig 1.

From the prediction algorithm, the classification tree produced 95.4% of accuracy. Similarly the accuracy for the other algorithms are given in table 2.

Table 2: Classification Algorithms and Accuracy

Methods used	Accuracy Percentage
Classification tree	95.4%
Naïve Bayes	81.7%
KNN	75.7%
Random Forest Classification	96.3%
SVM	100%

The confusion matrix generated in orange tool calculates the accuracy and specification measures. The matrix validates the effectiveness of the algorithms and finally concluded that support vector machine produced 100% accuracy for the prediction of disease.

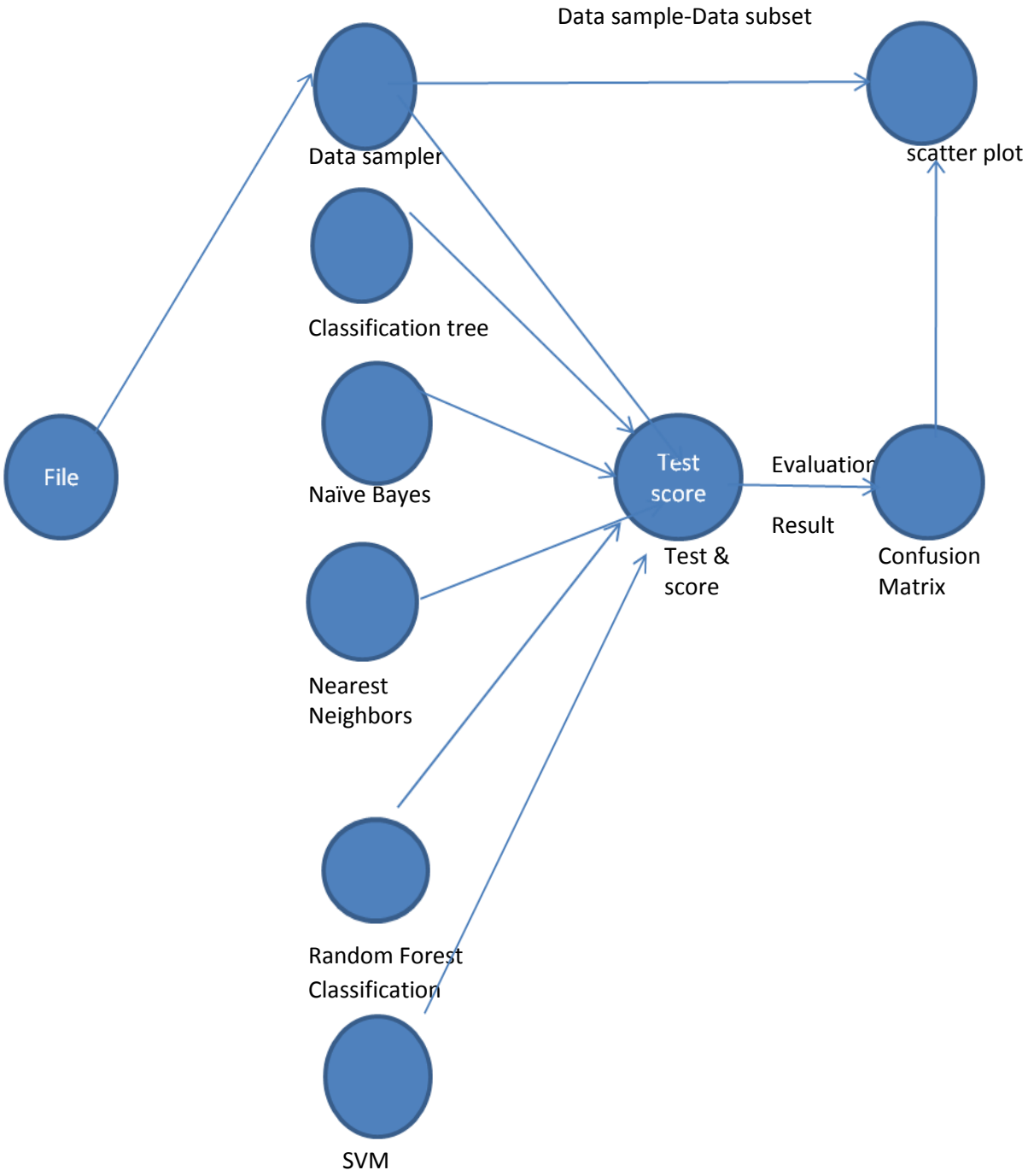

Fig 1: Classify widget

5. CONCLUSION

The paper discussed how different types of supervised learning algorithms are used to diagnosis the heart disease and also observed that how these algorithms performed better results by applying in Orange tool. The analysis shows that different technologies used in all the papers by taking different number of attributes. In some papers, it shows that ANN given the accuracy of 99.25% in prediction of heart disease [10]. On the other hand, it is also given that Decision tree has performed well with 99.2% accuracy in prediction of disease [2]. So, it is concluded that different algorithms produced different accuracy depends upon the attributes taken and tools used for implementation. Also support vector machine produced 100% result and can be effectively used in prediction of heart disease in patients in the proposed work.

Future work should be focused on attribute reduction and these attributes are reduced based on fuzzy logic concept. This work can be further extended by working with different heart related datasets from various health care organizations and also we can test the accuracy using real time datasets.

6. REFERENCES

[1] Resul Das, Ibrahim Turkoglu and Abdulkadir Sengur., Effective diagnosis of heart disease through neural network ensembles. *Expert System with Application*, pp 7675-7680, 2009.

[2] Anbarasi Masilamani, Anupriya and N.C.H.S.N. Iyengar., Enhanced prediction of Heart Disease with feature subset selection using Genetic Algorithm. *International Journal of Engineering and Science Invention*, Vol 2(10), pp 1-4, 2010.

[3] Ansari.A.Q. and Neeraj Kumar Gupta., Automated Diagnosis of coronary heart disease using Neuro-Fuzzy integrated system. *World Congress on Information and Communication Technology*, pp 1383-1388, 2011.

[4] Nidhi Bhatla and Kiran Jyoti., 2012. An analysis of heart disease prediction using different data mining techniques. *International Journal of Engineering Research & Technology*, Vol 1(8), pp 1-4, Oct 2012.

[5] T.John Peter and K.Somasundaram, Study and development of novel feature selection framework for heart disease prediction. *International Journal of Scientific and Research Publications*, Vol 2(10), 2012.

[6] Chaitrail S.Dangare and Sulabha S.Apte, Data mining approach for prediction of heart disease using neural network. *International Journal of Computer Engineering & Technology(IJCET)*, Vol 3(3), pp 30-40, Oct 2012.

[7] Jesmin Nahar, Tasadduq Imam and Kevin Tickle.S & Yi – Ping Chan., Computational Intelligence for heart disease diagnosis: A medical knowledge driven Approach. *Expert System with Application* Vol 40. pp 96-104, 2013.

[8] Thenmozhi.K and Deepika.P, Heart Disease Prediction using classification with different decision tree techniques. *International Journal of Engineering Research & General Science*, Vol 2(6), pp 6-11, Oct 2014.

[9] Deepali Chandna, Diagnosis of heart disease using data mining algorithm. *International Journal of Computer Science and Information Technology*, Vol 5(20), pp 1678-1680, 2014.

[10] Tamilarasi.R and Dr.Porkodi.R., A study and analysis of disease prediction techniques in data mining for healthcare. *International Journal of Emerging Research in Management and Technology*, Vol 4(3), pp 76-82, March 2015.

[11] Aravinthan.K and Dr.Vanitha.M., A comparative study on Prediction of Heart Disease using Cluster and Rank based Approach. *International Journal of Advanced Research in Computer and Communication Engineering*, Vol 5(2), pp 421-424, February 2016.

[12] Akhil jabbar.M, Dr.Priti Chandra and Dr.B.L. Deekshatulu., Heart disease prediction system using Associative classification and Genetic Algorithm. *International Conference on Electrical and Information Technologies*, Vol (1), pp 183-192, Dec 2012.

[13] Imran Kurt, Mevlut Ture and Turhan Kurum.A., Comparing performances of logistic regression, classification and regression tree and neural networks for predicting coronary artery disease. *Expert System with Application*, Vol 3(4), pp 366-374, 2008.

[14] Hlaudi Daniel Masethe and Mkosima Anna Masethe., Prediction of Heart Disease using Classification Algorithms. *World Conference on Electrical and Computer Science*, Vol II, pp 22-24, October 2014.

[15] Suganya.R, Rajaram.S, Sheik Abdullah.A and Rajendran.V., A Novel Feature Selection Method for Predicting Heart Disease with Data Mining Techniques. *Asian Journal of Information Technology*, Vol 15(8), pp 1314-1321, 2016.

[16] Beant Kaur and Williamjeet Singh., Review on Heart Disease Prediction system using Data Mining Techniques. *International Journal on Recent and Innovation Trends in Computing and Communication*, Vol 2(10), pp 3003-3008, 2014.

[17] Subanya.B and Rajalaxmi.R.R, A Novel Feature Selection Algorithm for Heart Disease Classification, *International Journal of Computation Intelligence and Informatics*, Vol 4(2), pp 117-124, 2014.

[18] Sagar Imambi.S and Sudha.T,, A Novel Features Selection Method of Classification of Medical Documents from Pubmed. *International Journal of Computer Applications*, Vol 26(9), pp 29-33, July 2011.