Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 85 (2016) 862 – 870

Procedia Computer Science

International Conference on Computational Modeling and Security (CMS 2016)

Analysis of Data Mining Techniques For Healthcare Decision Support System Using Liver Disorder Dataset

Tapas Ranjan Baitharu[a], Subhendu Kumar Pani[b*]

[a]*Orissa Engineering College,BPUT,Bhubaneswar,752050,India*
[b]*Orissa Engineering College,BPUT,Bhubaneswar,752050,India*

Abstract

Accuracy in data classification depends on the dataset used for learning. Now-a-days the most important cause of death for both men and women is due to the Liver Problem. The healthcare industry collects a huge amount of data which is not properly mined and not put to the optimum use. Discovery of these hidden patterns and relationships often goes unexploited. Our research focuses on this aspect of Medical diagnosis by learning pattern through the collected data of Liver disorder to develop intelligent medical decision support systems to help the physicians. In this paper, we propose the use decision trees J48, Naive Bayes, ANN, ZeroR, 1BK and VFI algorithm to classify these diseases and compare the effectiveness, correction rate among them. Detection of Liver disease in its early stage is the key of its cure. It leads to better performance of the classification models in terms of their predictive or descriptive accuracy, diminishing of computing time needed to build models as they learn faster, and better understanding of the models. In this paper, a comparative analysis of data classification accuracy using Liver disorder data in different scenarios is presented. The predictive performances of popular classifiers are compared quantitatively.

Peer-review under responsibility of the Organizing Committee of CMS 2016

Keywords: Classification; Data Mining;J48; Naive Bayes; Artificial Neural Network; 1BK; VFI;

Corresponding author. Tel.: 91-9776503280; fax: 06758-239723.
*E-mail address:*skpani.india@gmail.com

Peer-review under responsibility of the Organizing Committee of CMS 2016
doi:10.1016/j.procs.2016.05.276

1. Introduction

Data and information have become major assets for most of the organizations [1,4]. The success of any organization depends largely on the extent to which the data acquired from business operations is utilized. In other words, the data serves as an input into a strategic decision making process, which could put the business ahead of its competitors . Also, in this era, where businesses are driven by the customers, having a customer database would enable management in any organization to determine customer behavior and preference in order to offer better services and to prevent losing them resulting better business. In this research work, J48, Naive Bayes, ANN, ZeroR, 1BK and VFI algorithm classifier algorithms are used for liver disease prediction. There are various numbers of liver disorders that required clinical care of the physician [3]. The main objective of this research work is to forecast liver diseases such as Cirrhosis, Bile Duct, Chronic Hepatitis, Liver Cancer and Acute Hepatitis from Liver Function Test (LFT) dataset using above classification algorithms[2].

The liver is the second largest inside organ in the human body, playing a key role in metabolism and serving several imperative functions, e.g. Decomposition of red blood cells, etc.Its weight is around three pounds. The liver does many essential functions related to digestion, metabolism, immunity, and the storage of nutrients within the body. These functions formulate the liver as an important organ, without this, body tissues would rapidly die from lack of energy and nutrients. There are number of factors which boost the risk of liver disease.

Data mining is regarded as an emerging technology that has made radical change in the information world. The term `data mining' (often called as knowledge discovery) refers to the method of analyzing data from different perspectives and summarizing it into valuable information by means of a number of analytical tools and techniques, which in turn may be useful to increase the performance of a system. Technically, “data mining is the method of finding correlations or patterns among dozens of fields in large relational databases”. Therefore, data mining consists of key functional elements that transform data onto data warehouse, manage data in a multidimensional database, facilitates data access to information professionals or analysts, analyze data using application tools and techniques, and meaningfully present data to provide useful information [5,6].

The paper is divided into five sections. Section 2 describes the main techniques and algorithms associated with data mining. In Section 3, we provide the results. We provide descriptive results. In this section we also compare the predictive power of the classifiers. Finally, in Section 4, we draw some conclusions from our results and offer some directions for future research.

2. Techniques and Algorithms

Researchers find two important goals of data mining: prediction and description. First, the Prediction is possible by use of existing variables in the database in order to predict unknown or future values of interest. Second the description mainly focuses on finding patterns describing the data the subsequent presentation for user interpretation. The relative emphasis of both prediction and description differs with respect to the underlying application and technique.

Classification: Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier. Some well-known classification models are

2.1 .Decision trees J48

J48 [14] is an important decision tree classifier. Decision tree is a predictive machine-learning representation that makes decisions the target value (dependent variable) of a fresh sample based on various attribute values of the available data. The internal nodes of a decision tree indicate the different attributes, the branches between the nodes gives the probable values that these attributes can have in the observed samples, while the terminal nodes generates the final value (classification) of the dependent variable. The attribute that is to be predicted is recognized as the dependent variable, since its value depends upon, or is chosen by, the values of all the other attributes. The other attributes, which help in predicting the value of the dependent variable, are known as the independent variables in the dataset.Figure1 shows the decision tree J48 implementation using Liver Disorder Dataset.

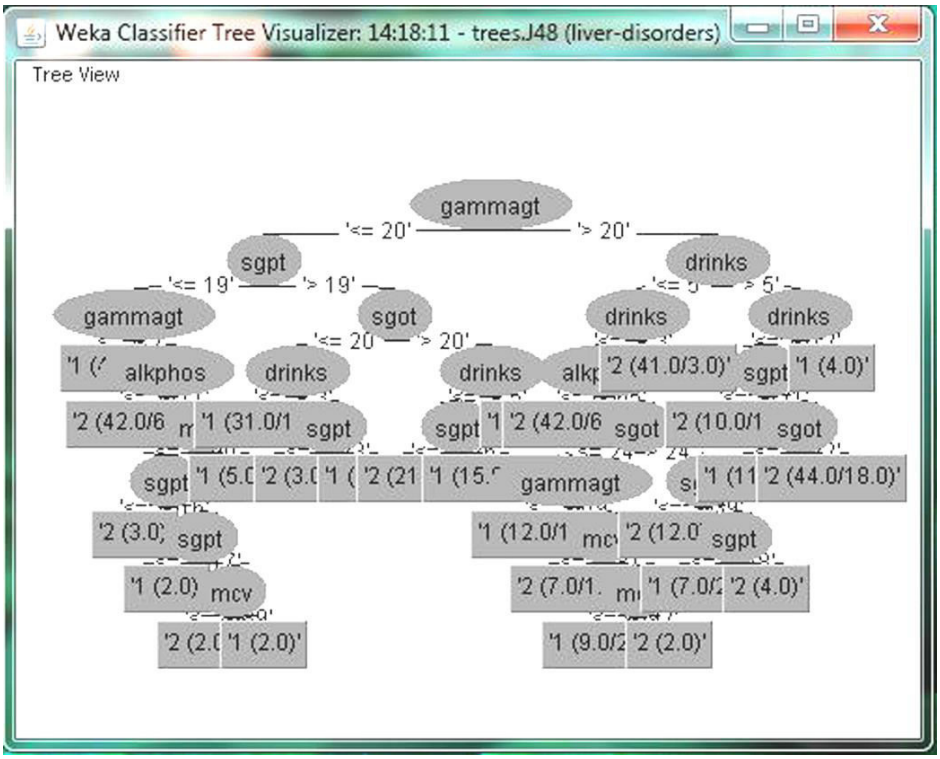

Figure 1 : J48 Tree visualizer

2.2. Naive Bayes

The Naive Bayesian classifier is basically on Bayes' theorem with independence assumptions between predictors [13]. A Naïve Bayesian model is very simple to build and can be implemented for very huge datasets. Naive Bayesian classifier often achieves well than more sophisticated classification techniques. The posterior probability, P (c|x) is computed from P(c), P(x), and P (x|c). 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 named class conditional independence [9,10].

2.3. Multilayer perceptron

A multilayer perceptron (MLP) is a feed forward artificial neural network model that charts sets of input data onto a set of appropriate outputs. An MLP composes of multiple layers of nodes in a directed graph, with every layer fully

connected to the next one. Except for the input nodes, every node is a neuron (or processing element) with a nonlinear activation function. MLP uses a supervised learning technique called back propagation for training the network. MLP is a alteration of the standard linear perception and can differentiate data that are not linearly separable [11, 12].
If a multilayer perceptron has a linear activation function in all neurons, that is, a linear function that charts the weighted inputs to the output of each neuron, then it is simply proved with linear algebra that any number of layers can be reduced to the standard two-layer input-output model. What makes a multilayer perceptron different is that some neurons use a nonlinear activation function which was developed to model the frequency of action potentials, or firing, of biological neurons in the brain. This function is modeled in several ways.

This function is modelled in several ways.

2.4 .ZeroR

ZeroR is the easiest classification method which depends on the target and ignores all predictors. ZeroR classifier basically predicts the majority category (class). Even though there is no predictability power in ZeroR, it is helpful for determining a baseline performance as a benchmark for other classification methods. A frequency table is created for the target and the most frequent value is selected.

2.5 .1BK

KNN (K- Nearest Neighbor) classifier alternatively known as IBK, a supervised learning algorithm, where a given data set is divided into a user specified number of clusters. Predict the same class as the adjacent instance in the training set. Training stage of the classifier keeps the features and the class label of the training sets. New objects are classified based on the voting criteria. It gives the maximum likelihood estimation of the class. Euclidean distance metrics is applied for assigning objects to the most frequently labeled class. Distances are computed from all training objects to test object using appropriate K value. It constructs the decision tree from labeled training data set using information gain and it observes the same that results from choosing an attribute for splitting the data. To build the decision the attribute with maximum normalized information gain is used. Then the algorithm recurs on smaller subsets. The splitting procedure ends if all instances in a subset belong to the same class. Then the leaf node is built in a decision tree telling to choose that class.

2.6. VFI algorithm

It is a very easy algorithm. Let us suppose there are 'm' no. of features and 'n' no. of classes involved. As every other classification algorithm, this classifier predicts a class as its final output. In this technique, each feature participates in the classification. Each feature presents a vote for one of the classes out of the n available classes. So for each class there are a total of m votes(1/feature).The class with the highest votes is affirmed to be the predicted class.
VFI is a unexpectedly very fast algorithm and performs very well in most of the cases; even better than the rest of the classifiers in some cases.

2.7.Margin Curve

The margin is described as the difference between the probability predicted for the actual class and the highest probability predicted for the other classes. One hypothesis as to the fine performance of boosting algorithms is that they boost the margins on the training data and this gives better performance on test data. The margin curves for J48,ZeroR, Multilayer Perceptron, 1BK, Naive Bayes, VFI are shown in figure2,figure-3,figure-4,figure-5 , figure-6,figure-7.

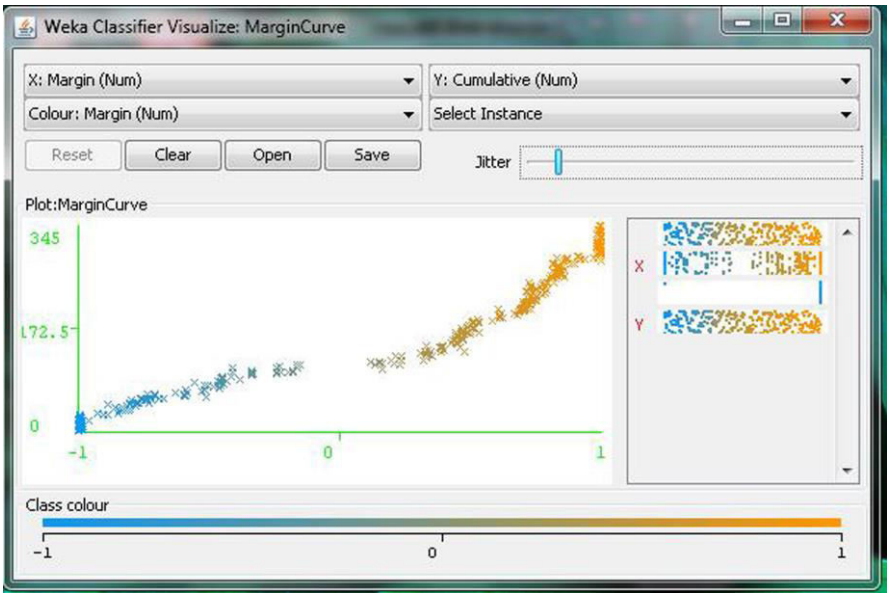

Figure2: Margin Curve of J48

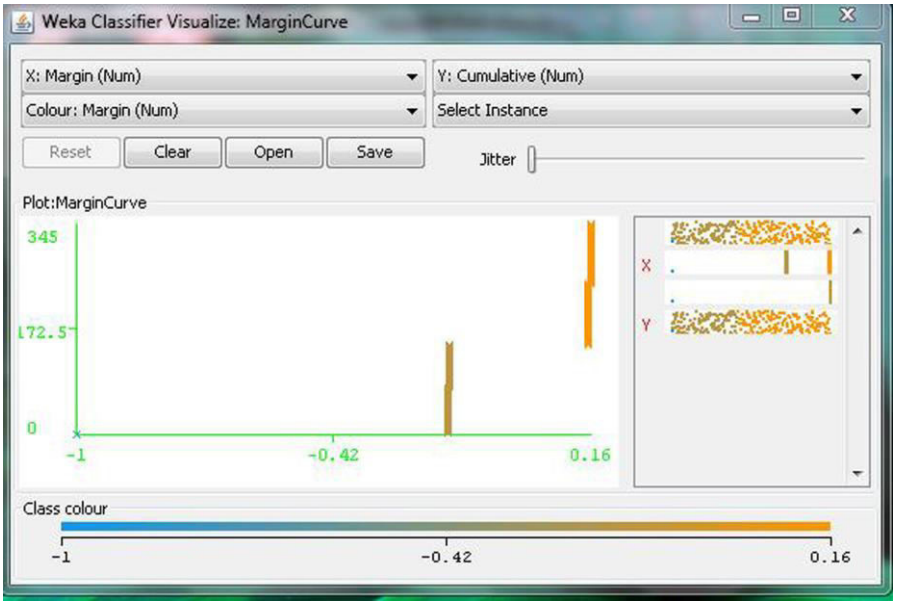

Figure3 :Margin Curve of ZeroR

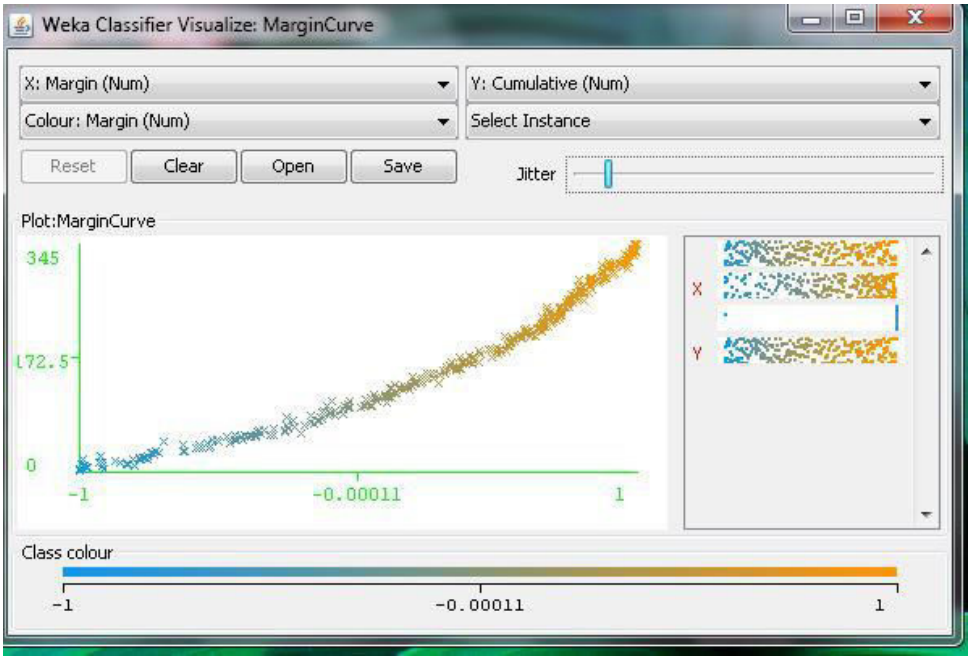

Figure 4:Margin Curve of Multilayer Perceptron

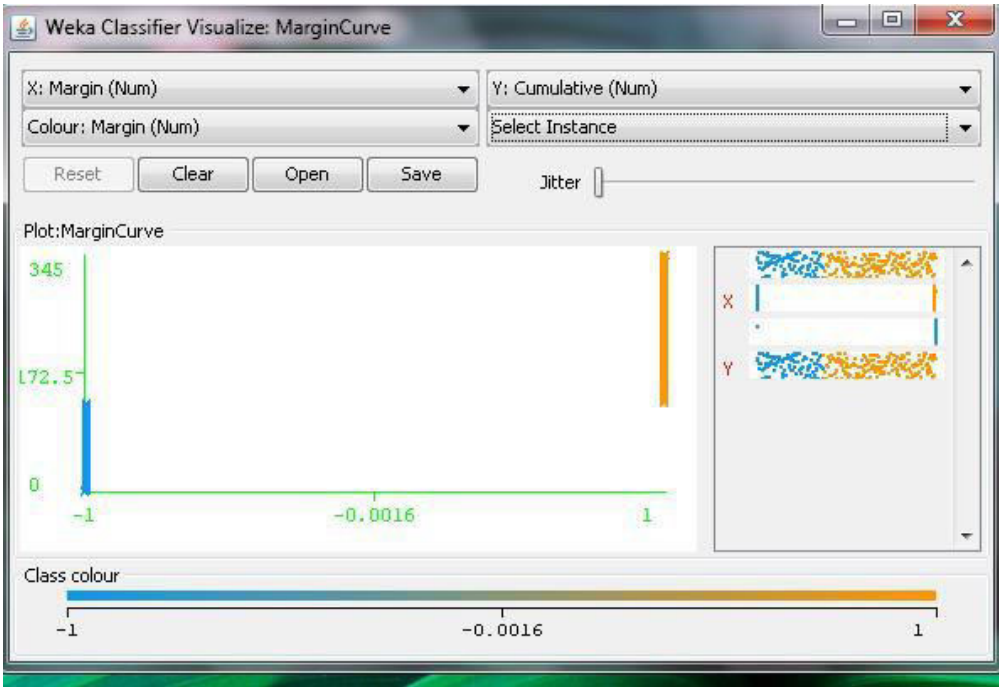

Figure5:Margin Curve of 1BK

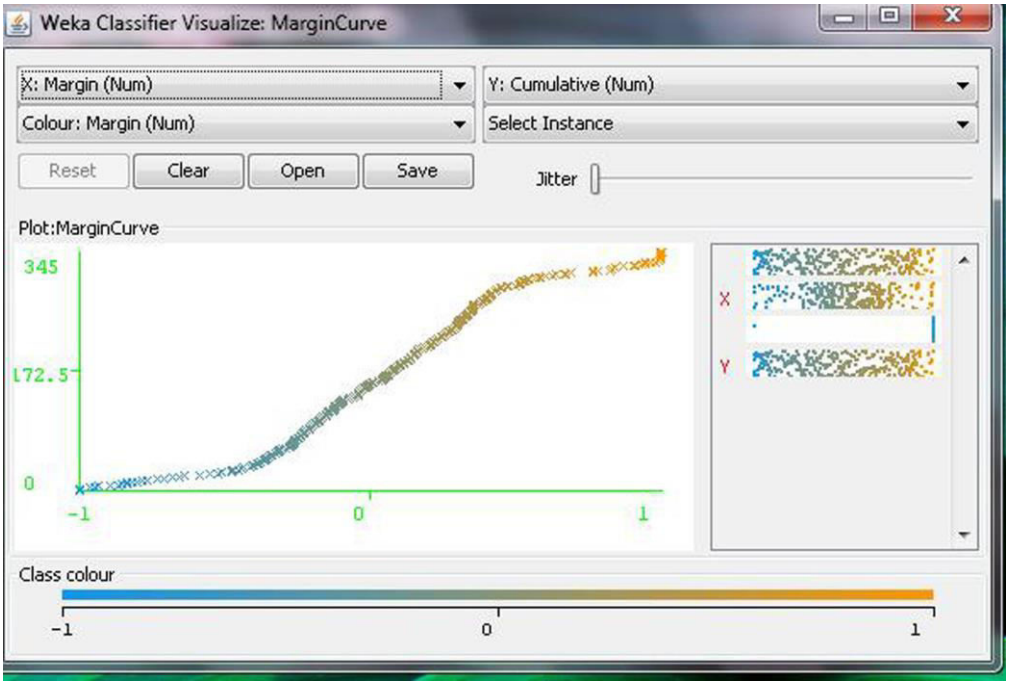

Figure 6 :Margin Curve of Naive Bayes

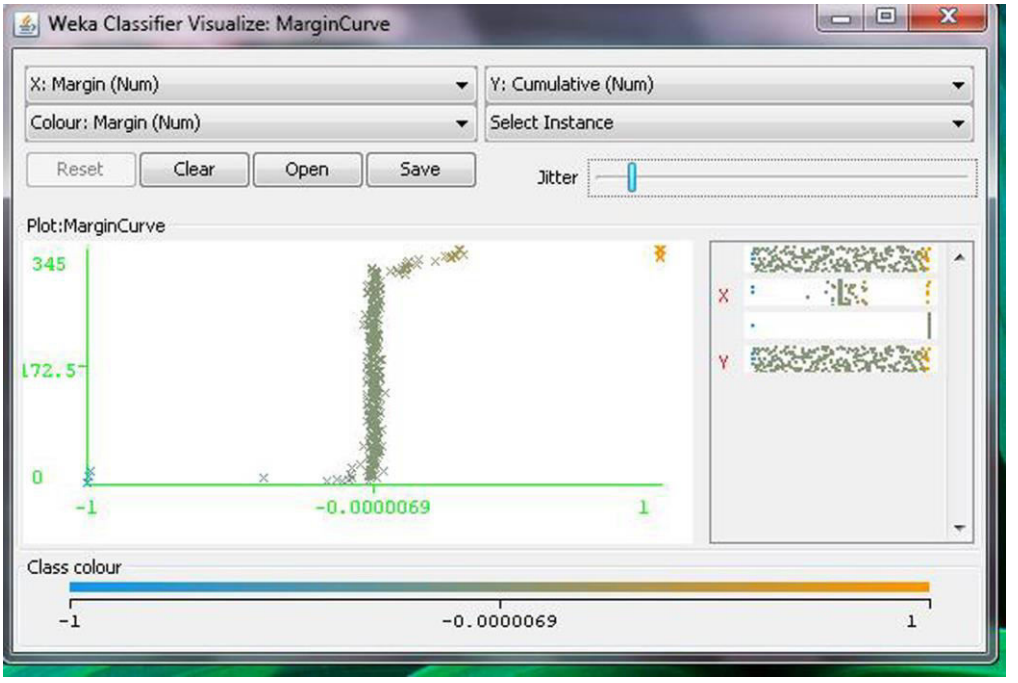

Figure7 :Margin Curve of VFI

3.Experimental Study and Analysis

3.1 WEKA Tool

We use WEKA (www.cs.waikato.ac.nz/ml/weka/), an open source data mining tool for our experiment. WEKA is developed by the University of Waikato in New Zealand that implements data mining algorithms using the JAVA language. WEKA is a state-of-the-art tool for developing machine learning (ML) techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. WEKA implements algorithms for data pre-processing, feature reduction, classification, regression, clustering, and association rules. It also includes visualization tools. The new machine learning algorithms can be used with it and existing algorithms can also be extended with this tool.

3.2 Classifier Selection

We select six commonly used classifier for prediction classification in our work based on their qualitative performance. These classifiers are briefly described and their performance is analyzed in are given below in Table 1.

Table 1. Performance analysis of different Classifiers using Liver Disorder Dataset.

Data Mining Algorithms	Kappa Statistics	Mean absolute Error	Root Mean squared Error	Relative absolute error	Accuracy
J48	0.3401	0.3673	0.5025	75.3511	68.97
ZeroR	0	0.4874	0.4936	100	57.971
Multilayer Perceptron	0.4023	0.3543	0.4523	72.68	71.59
1BK	0.2401	0.3718	0.6072	76.2906	62.8986
Naive Bayes	0.153	0.4597	0.5083	102.9673	55.3623
VFI	0.1044	0.4889	0.4955	100.3839	60.2899

3.3 Results Analysis

We run selected classifiers in different scenarios of the dataset. By analyzing the results, Multilayer perceptron gives the overall best classification result than other classifiers. : Accuracy Comparison of different Classifiers is shown in figure-8.

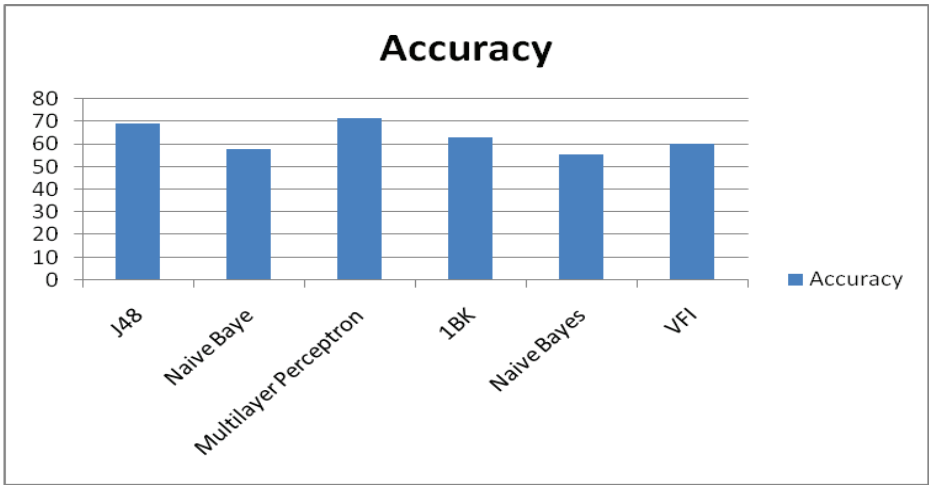

Figure8: Accuracy Comaprision of different Classifiers

4. Conclusion

An experiment is conducted to get the impact of liver disorder on the predictive performance of different classifiers. We select six popular classifiers considering their qualitative performance for the experiment. After analyzing the

quantitative data generated from the computer simulations, we find that the general concept of improved predictive performance of all above classifiers but Naive Bayes performance is not significant. However more experiments with different datasets are required to support the findings. Classification is the major data mining technique which is primarily used in healthcare sectors for medical diagnosis and predicting diseases. This research work used classification algorithms for liver disease prediction. Comparisons of these algorithms are done and it is based on the performance factors classification accuracy and execution time.

References

1. Klosgen W and Zytkow J M (eds.), Handbook of data mining and knowledge discovery, OUP, Oxford, 2002.

2. Provost, F., & Fawcett, T., Robust Classification for Imprecise Environments. Machine Learning, Vol. 42, No.3, pp.203-231, 2001.

3. Larose D T, Discovering knowledge in data: an introduction to data mining, John Wiley, New York, 2005.

4. Kantardzic M, Data mining: concepts, models, methods, and algorithms, John Wiley, New Jersey, 2003.

5. Goldschmidt P S, Compliance monitoring for anomaly detection, Patent no. US 6983266 B1, issue date January 3, 2006, Available at: www.freepatentsonline.com/6983266.html

6. Bace R, Intrusion Detection, Macmillan Technical Publishing, 2000.

7. Smyth P, Breaking out of the BlackBox: research challenges in data mining, Paper presented at the Sixth Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD2001), held on May 20 (2001), Santra Barbara, California, USA.

8. Agrawal R. and Srikant R. Fast Algorithms for Mining Association Rules. In M. Jarke J. Bocca and C. Zaniolo, editors, Proceeedings of the 20th International Conference on Verylarge DataBases (VLDB'94), pages 475–486, Santiago de Chile, Chile, Sept 1994 . MK

9. J. Su, H. Zhang, C.X. Ling, S. Matwin, Discriminative parameter learning for Bayesian networks, in: Proceedings of the Twenty-Fifth International Conference on Machine Learning, ACM Press, Helsinki, Finland, 2008, pp. 1016–1023.

10.A.A. Balamurugan, R. Rajaram, S. Pramala, et al., NB+: An improved Naïve Bayesian algorithm, Knowledge-Based Systems 24 (5) (2011) 563–569.

11. Sta´nczyk, U.. Establishing relevance of characteristic features for authorship attribution with ANN. In: Decker, H., Lhotska, L., Link,S., Basl, J., Tjoa, A., editors. Database and Expert Systems ppllications; vol. 8056 of Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2013, p. 1–8.

12. Sta´nczyk, U.. Rough set and artificial neural network approach to computational stylistics. In: Ramanna, S., Jain, L.C., Howlett, R.J.,editors. Emerging Paradigms in Machine Learning; vol. 13 of Smart Innovation, Systems and Technologies. Springer Berlin Heidelberg;2013, p. 441–470.

13.Sta´nczyk, U.. Decision rule length as a basis for evaluation of attribute relevance. Journal of Intelligent and Fuzzy Systems 2013;24(3):429–445.

14. Farid, D.M., Zhang, L., Rahman, C.M., Hossain, M., Strachan, R.. Hybrid decision tree and Naive Bayes classifiers for multi-class classification tasks. Expert Systems with Applications 2014; 41(4, Part 2):1937–1946.