

## **Comparative Analysis of Classification Techniques using WEKA on Different Datasets**

**Mahesh Parmar**

*Department of CSE & IT, M.I.T.S. Gwalior, India*

---

**Abstract:** Data mining is the analytic process designed to explore large amounts of data in search for consistent patterns and systematic relationships between variables and then to validate the findings by applying the detected patterns to new subsets of data. Data mining software are analytical tools for analyzing data. Weka is a data mining tools, contains many machine learning algorithms and provides the facility to classify our data through various algorithms. Classification techniques a model is built based on training data and applied to test data in broad applications. In this paper, two classification algorithms are used for analyzing datasets. The main aims to show the comparative Analysis of decision tree (J48) and Backpropagation classification algorithm using WEKA tool and find out which technique is most suitable for user working on different datasets. The best algorithm based on the Bank datasets and Vote dataset is MLP classifier with accuracy respectively of 73.75. % and 96.32%.

**Keywords:** Classification, Data Mining Techniques, Decision Tree, Multilayer Perceptron

---

### **I. INTRODUCTION**

The past decade has seen an explosive growth in database technology and the amount of data collected. Advances in data collection, use of bar codes in commercial outlets, and the computerization of business transactions have flooded us with lots of data. We have an unprecedented opportunity to analyze this data to extract more intelligent and useful information, and to discover interesting, useful, and previously unknown patterns from data.

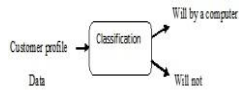
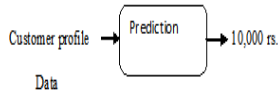
Data mining technique [1] is a process of discovering pattern of data. The patterns discovered must be meaningful in that they lead to some advantage. In recent times, data mining has been obtained a great attention in the knowledge and information industry due to the vast availability of large amounts of data and the forthcoming need for converting such data into meaningful information and knowledge. Data mining is being used in several applications like banking, insurance, and hospital and Health informatics [2] [3]. In case of health informatics, Data mining plays a vital role in helping physicians to identify effective treatments, and Patients to receive better and more affordable health services. In hematology laboratory, it has become a powerful tool in managing uncountable laboratory information in order to seek knowledge that is underlying or within any given information.

The aim of data mining is to extract implicit, previously unknown and potentially useful patterns from data. Data mining consists of many up-to-date techniques [4] such as classification, clustering, and association [5]. Many years of practice show that data mining is a process, and its successful application requires data preprocessing, post processing, good understanding of problem domains and domain expertise

#### **A. Classification and Prediction**

There are two forms of data analysis, classification and prediction that can be used for extracting models describing important classes or to predict future data trends. Classification models predict categorical class labels and prediction models predict continuous valued functions. We can build a classification model to categorize bank loan applications as either safe or risky, or a prediction model to predict the expenditures in dollars of potential customers on computer equipment given their income and occupation [6].

**B. Differences between Classification and prediction**

S. No.	Classification	prediction
1.	Predicts categorical class labels (discrete or nominal)	Models continuous-valued functions, i.e., predicts unknown or missing values
2.	<p>Example: A model or classifier is constructed to predict categorical labels such as “safe” or “risky” for a loan application data.</p> 	<p>Example: A marketing manager would like to predict how much a given customer will spend during a sale</p> 

There are two main steps in classification. Step1: Model Construction: Construct a classification model based on training data, Training data a set of tuples, Each tuple is assumed to belong to a predefined class with labeled data. Step2: Model Usage: If the accuracy is acceptable, use the model to classify data tuples whose class labels are not known. Before using the model, we first need to test its accuracy

**II. CLASSIFICATION ALGORITHMS****A. Decision Tree**

It is a well known classification method [8] that takes the form of tree structure and it is usually made up of:

- 1) Testing node which holds the data for testing the condition.
- 2) Start node is the parent and usually top most node.
- 3) Terminal node (leaf node): is the predicted class label
- 4) Branches: represents results of a test made on an attribute.

**Decision Tree Algorithm****Parameters**

Datase T and its fields

Set of Attributes A

Selection Technique for the Attribute

**Procedure**

- A node is Created (call it E)
- Check if all records R is in one group G and write node E as the last node in the that Group G
- If  $A=0$  ( no attribute)
- then write E as the last node
- Use Selection technique for attributes on (R, A) to get the Best splitting condition
- Write the condition on node E
- Check if attribute is discrete and allows multiway split then It is not strictly binary tree
- For all output O from splitting condition, divide the records and build the tree
- Assign  $R_0$  = Set of all records in output O
- If  $R_0 = 0$  then
- Node E is attached with a leaf labelled with majority class R
- Otherwise node E is attached with node obtained from Generate Decision Tree ( $R_0$ , A)
- Next
- Write E

**Result**

Tree Classifier

**B. Back propagation Algorithm**

Back-propagation [14] training algorithm when applied to a feed forward multi-layer neural network is known as Back propagation neural network. Functional signals flows in forward direction and error signals propagate in backward direction. That's why it is Error Back Propagation or shortly Back Propagation network. The activation function [13] that can be differentiated (such as sigmoid activation function) is chosen for hidden and output layer computational neurons. The algorithm is based on error-correction rule. The rule for changing values of synaptic weights follows generalized delta rule.

**Steps:**

*Initialize all weights in network*  
*//Propagate the inputs forward*

- For each input layer unit  $j$
- $O_j = I_j$  //output of an input unit its actual input value
- For each hidden or output layer unit  $j$
- $I_j = \sum_i w_{ij} O_i$  // the net input of unit  $j$
- $O_j = 1/(1+e^{-I_j})$  // the output of each unit  $j$
- //Back propagate the errors
- For each unit  $j$  in the output layer
- $Err_j = O_j (1-O_j)(T_j - O_j)$  // the error
- For each unit  $j$  in the hidden layer, from the last to the 1<sup>st</sup> hidden layer
- $Err_j = O_j (1 - O_j) \sum_k Err_k w_{jk}$  //error with respect to the
- Next higher layer ,  $k$
- for each weight  $w_{ij}$  in network
- $\Delta w_{ij} = (\eta) Err_j O_i$  //weight increment
- $W_{ij} = w_{ij} + \Delta w_{ij}$  //weight update

### III. EXPERIMENTAL DETAILS

#### A. Datasets

There are two datasets we have used in our paper taken from UCI Machine Learning Repository [12]. Bank Dataset : In Bank dataset there are 11 attributes (age, sex, region, income, married, children, car, save-account, current account, mortgage and pep) and 600 data items, Classified into two classes, the classification is done whether the person will go for Pension Equity Plan (PEP) or not. The details of each datasets are shown in Table 1

Vote Dataset: In this dataset we have 16 attributes and 435 data instances, classified into one class, the class have two values democrat or republican.

Table 1: Datasets

Datasets	Instances	Attributes	No. of Classes	Type
Bank Dataset	600	11	2	Multivariate
Vote Dataset	435	16	1	Multivariate

The database connectivity was established with Weka Tool for further analysis by applying data mining technique. Different parameters were set before applying technique.

#### B. Experimental Result

A comparison of classifiers for different datasets, The confusion matrix helps us to find the various evaluation measures like Accuracy, Recall, Precision etc the accuracy and time taken for execution is made. Accuracy is defined as the no. of instances classified correctly.

Table 2: Evaluation parameters on Bank dataset

S. No.	Parameters	MLP	J48
1	TP Rate	0.817	0.738
2	FP Rate	0.062	0.088
3	Precision	0.814	0.731
4	Recall	0.817	0.738
5	F-Measure	0.815	0.733
6	ROC Area	0.948	0.898

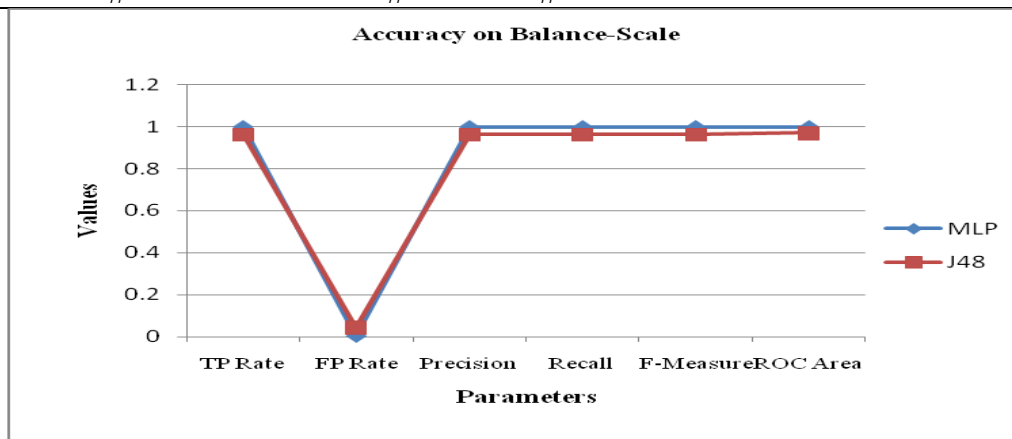


Fig. 1: Accuracy graph of Bank dataset

In Bank dataset evaluation parameters have shown in Table 2 and Fig 1. The above chart shows that Classification algorithm J48 having lower parameters value except FP Rate as compare to MLP. It is interpreted that MLP is better method for Bank dataset.

Table 3: Evaluation parameters on Vote dataset

S. No.	Parameters	MLP	J48
1	TP Rate	0.998	0.963
2	FP Rate	0.004	0.041
3	Precision	0.998	0.963
4	Recall	0.998	0.963
5	F-Measure	0.998	0.963
6	ROC Area	0.998	0.971

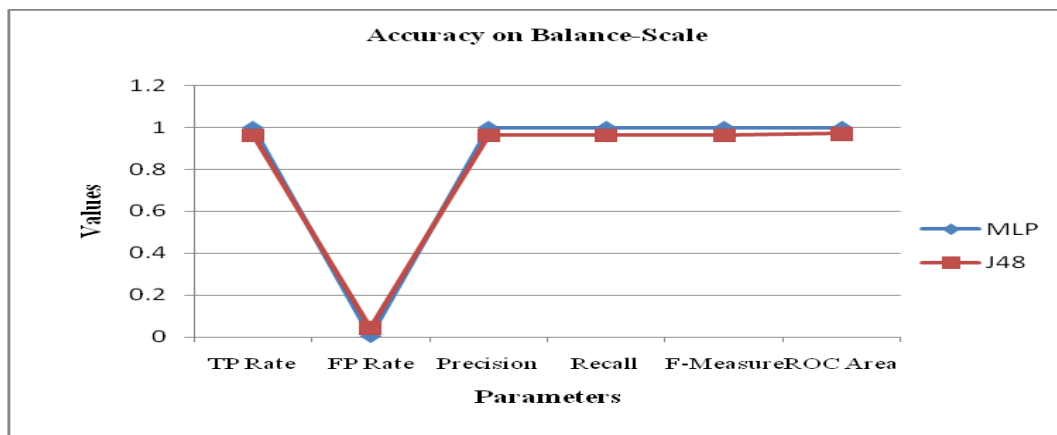


Fig. 2: Accuracy graph of Vote dataset

In vote dataset accuracy parameters have shown in Table 3 and Fig 2. The above chart shows that MLP classification algorithm has almost equal accuracy measures except FP rate as compare to J48 classification algorithm. So, MLP is better method for vote dataset.

Table 4: Comparative accuracy on Datasets

Parameters	MLP	J48
Bank Dataset	81.67	73.75
Vote Dataset	99.77	96.32

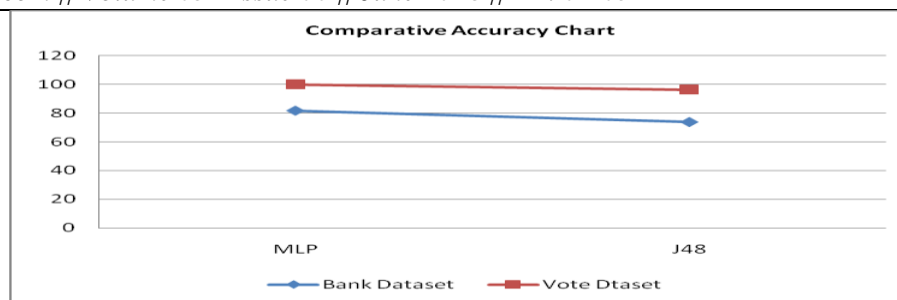


Fig. 3: Comparative accuracy on datasets

The J48 and MLP classification algorithm applies on both the datasets for accuracy measure. From the values of Table 4 and the chart shown in Fig 3, the accuracy measures are calculated on J48 and MLP algorithms. It is clear that MLP produces better results for both datasets so that MLP is better algorithm than J48 for the given datasets

#### IV. CONCLUSION

In this paper we have studied and compared Decision Trees (J48) and MLP classification algorithm on two data sets in WEKA. We evaluate the performance in terms of classification accuracy of J48 and Multilayer Perceptron algorithms using various accuracy measures like TP rate, FP rate, Precision, Recall, F-measure and ROC Area. Accuracy has been measured on each datasets. Overall observation is that the best algorithm based on the both datasets is MLP. Generally neural networks have not been suited for data mining but from the above results we conclude that algorithm based on neural network has better learning capability hence suited for classification problems if learned properly.

#### REFERENCES

- [1] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann, 2nd, 2006
- [2] Z. Haiyang, "A Short Introduction to Data Mining and Its Applications", IEEE, 2011
- [3] Kritika Yadav and Mahesh Parmar: Analysis of Mahatma Gandhi National Rural Employment Guarantee Act Using Data Mining Technique. International Journal of Computational Intelligence Research (IJCIR). 2017, Volume 13, Number 9 (2017), pp. 2221-2235, ISSN 0973-1873
- [4] Kritika Yadav, Mahesh Parmar: Review Paper on Data Mining and its Techniques and Mahatma Gandhi National Rural Employment Guarantee Act. International Journal of Computer Science and Engineering (JCSE), April 2017, Volume-5, Issue-4, pp. 68-73, E-ISDN: 2347-2693.
- [5] Moksha Shridhar, Mahesh Parmar: Survey on Association Rule Mining and Its Application. International Journal of Computer Science and Engineering (JCSE), March 2017, Volume-5, Issue-3, pp. 129-135, E-ISDN: 2347-2693.
- [6] Saichanma, Sarawut, Sucha Chulsomlee, Nonthaya Thangrua, Pornsuri Pongsuchart, and Duangmanee Sanmun. "The Observation Report of Red Blood Cell Morphology in Thailand Teenager by Using Data Mining Technique." Advances in hematology 2014.
- [7] Y. Freund and L. Mason. The alternating decision tree algorithm. In Proceedings of the 16th International Conference on Machine Learning, pages 124-133, 1999
- [8] Sumit Garg , AK. Sharma, "Comparative Analysis of Data Mining Techniques on Educational Dataset", International Journal of Computer Applications, Vol.74, No.5, pp.1-7, 2013.
- [9] Bhavesh Patankar, Vijay Chavda, "A Comparative Study of Decision Tree, Naïve Bayesian and k-nn Classifiers in Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Vol.4, Issue.12, pp.24-31, 2014.
- [10] kumar Y, Sahoo G. Analysis of Bayes, Neural Network and Tree Classifier of Classification Technique in Data Mining using WEKA Computer Science and Information Technology (CS & IT)-CSCP. 2012.
- [11] David, Satish Kumar, Amr TM Saeb, and Khalid Al Rubeaan. "Comparative Analysis of Data Mining Tools and Classification Techniques using WEKA in Medical Bioinformatics." Computer Engineering and Intelligent Systems 4, no. 13 (2013): 28-38.
- [12] Anshul Goyal, Rajni Mehta, "Performance Comparison of Naive Bayes and J48 Classification Algorithms", IJAER, Vol. 7, No. 11, 2012, pp.
- [13] Fisher, A. W., Fujimoto, R. J. and Smithson, R. C.A. (1991) 'A Programmable Analog Neural Network Processor', IEEE Transactions on Neural Networks, Vol. 2, No. 2, pp. 222-229.
- [14] Pai, G. V and Rajasekaran, S, (2006), 'Neural Networks, Fuzzy Logic and Genetic Algorithms Synthesis and Applications', 6th ed, Prentice Hall of India Pvt. Ltd.