ISSN:0975-9646

Suman et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3448-3453

A Comparative Performance Analysis of Classification Algorithms Using Weka Tool Of Data Mining Techniques

Suman [#1,] Mrs.Pooja Mittal[*2]

[#1]*Student of Masters of Technology, Department of Computer Science and Application*
M.D. University, Rohtak, Haryana, India

[*2]*Assistant Professor, Department of Computer Science and Application*
M.D. University, Rohtak, Haryana, India

Abstract:-We are live in a time there we used a huge amount and useful information and we want to save the time for this reason we use the data mining. Today we are using the electronic devices to store the information they have ability to save the very large amount of information so that to operate the information manually is very difficult and time consuming thus we use the data mining. Data mining describes a collection of techniques that aim to find useful, but undiscovered patterns in collecting data. Classification is one of them technique of data mining, which used to predefine classification data. Data mining provides many software to analysis the techniques. Weka is a tool which has allowed the users to analysis the data. In this paper, we are analysis various classification methods (classification by decision tree, Bayesian classification, neural network) and compare them using weka then we provide that which methods is better for users.

Keywords: - data mining, classification, weka tool and classification methods etc.

1. INTRODUCTION

Their various methods of data mining have been used in both, commercial and research centers. These methods can be used in education and industrial sectors to improve their performance. As we know that data mining can be applied to large database to explore the hidden information and describe the classification and pattern of data sets. Raw data is useless without techniques to extract Information from it. According to Data Mining, by I.H. Witten and E. Frank, "Data mining is defined as the process of discovering patterns in data. The process must be automatic or (more usually) semiautomatic. Knowledge Discovery in Databases (KDD) has emerged technology to analysis the huge amount of data.

2. CLASSIFICATION

Classifying data into a fixed number of groups (Soman et al., 2006) and using it for categorical variables (Nisbet, 2009) is known as classification [1]. Classification is divided into two types one is supervised and unsupervised. When an object is already known about its class is known as supervised if an object is not known about its class is known as unsupervised. Fraud detection and credit risk applications are particularly well suited to this type of analysis. In data classification we use learning and classification. Classification is classified into different models, these are followed:-

Types of classification models:-

- Classification by decision tree induction
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Associations

3. WEKA TOOL

Weka is a landmark system in the history of the data mining and machine learning, research communities, because it is the only toolkit that has gained such widespread adoption and survived for an extended period of time [2]. Its name is based on an endemic bird of New Zealand. Weka is open source tool and freely available. Weka tool is mainly used to analyze the data mining algorithm. Weka tool provides many algorithms for data mining and machine learning. Weka is platform-independent software. These tools and software provide a set of methods and algorithms that help in better utilization of data and information available to users; including methods and algorithms for data analysis, cluster analysis, Genetic algorithms, Nearest neighbor, data visualization, regression analysis, Decision trees, Predictive analytics, Text mining, etc. [7]

Figure3. 1View of weka tool.

Weka GUI chooser consists four buttons these are:-

- Explorer: - Explorer: An environment for exploring data.
- Experimenter: An environment for performing experiments and conducting statistical tests between learning schemes.
- Knowledge Flow: This environment supports essentially the same functions as the Explorer, but with a drag and- drop interface. One advantage is that it supports incremental learning.
- Simple CLI: Provides a simple command-line interface that allows direct execution of WEKA commands from operating systems that do not provide their own command line interface.

4. METHODOLOGY

In this paper, we compare the various classification techniques and provide the result for this purpose, we need the data set. So that we are using the educational data set which is mainly related to students.

5. PERFORMING CLASSIFICATION ON WEKA

We are performing classification on weak tool for that we are loaded the weak tool shown in fig5.1.The data should in the format of arff and csv because weak use the formats. Here we use the csv format database In this database instance, are 624 and attribute are16.

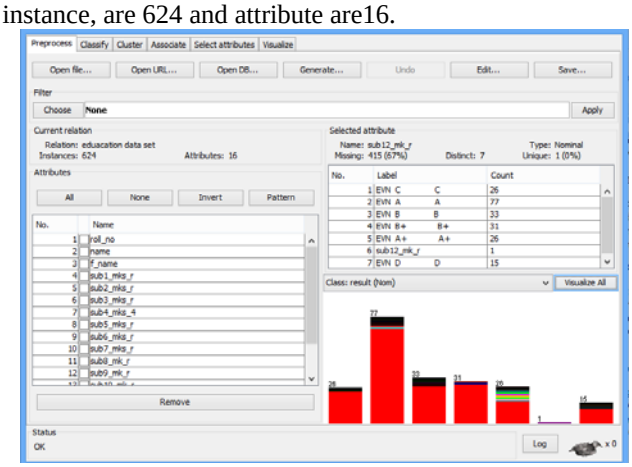
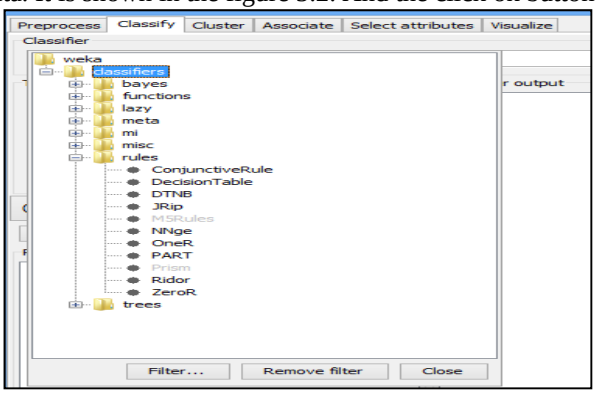


Fig5.1load data set in the wake.

We have many options shown in the figure5. 1. We perform classification so we click on the classify button. After that we choose an algorithm which is applied to the data. It is shown in the figure 5.2. And the click ok button



F igure5.2: various Classification algorithms in weka

6. BAYESIAN NETWORKS

A Bayesian Network (BN) is a graphical model which is used to provide relationships among a set of variable features. [3]. fig6.1. show the structure of Bayesian net.

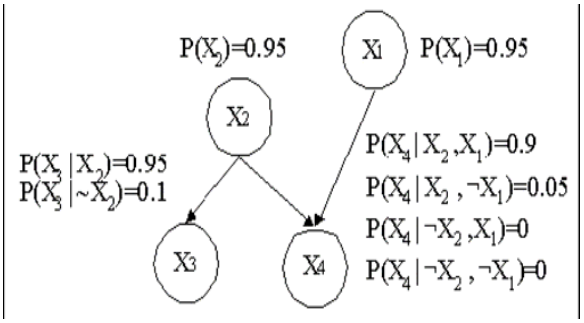


Fig6.1 basic structure of Bayes net

A. *Accuracy*: The measure of the Accuracy of the education dataset for BayesNet classifier technique is shown below with graph according to the Table No. 6.1From the below Table No.6.1 shows the training size (%), total no. instances, correctly classified instances and Kappa Statistic and Figure No.6.2 show the performance of Accuracy on education dataset. From the below Figure No.6.2 we can clearly see that the highest accuracy is 28.9773 % and lowest is 9.2593 % when the training size is 40% and 80% respectively. The accuracy decreases from 28% to 9%. The accuracy sometimes increases and sometimes decreases on the different types of training size. Here we can clearly see that the accuracy is increased when dataset is small split and when dataset is large split the accuracy is minimized.

Training Size (%)	Total no. of Instances (624)	Correctly Classified Instances % (value)	Incorrectly Classified Instances % (value)	Mean Absolute Error	Kappa Statistic
40	176	28.9773 %	71.0227 %	0.0344	-0.0187
50	147	23.1293 %	76.8707 %	0.4986	0.0273
60	118	16.1017 %	83.8983 %	0.0358	0.0237
70	82	17.0732 %	82.9268 %	0.1371	0.0103
80	54	9.2593 %	90.7407 %	0.0374	-0.0023

Table 6. 1. Show the result of Bayes net

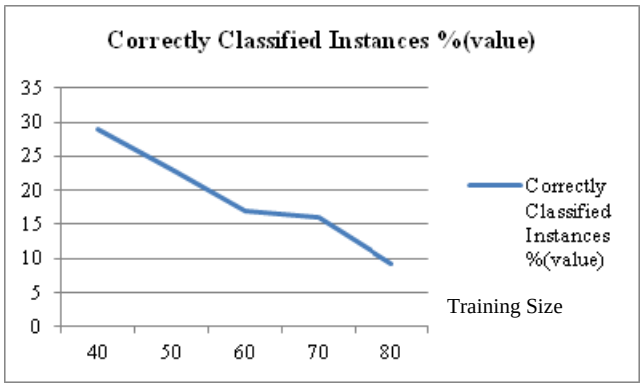


FIG 6.2 show the performance of Accuracy.

B. *Kappa Statistic*: The measure of the Kappa Statistic of the mushroom dataset for BayesNet classifier techniques is shown below with graph according to the Figure No.6.3.

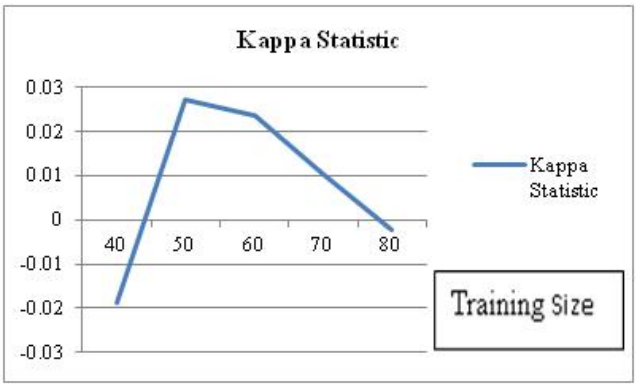


Figure No. 6.3 Show the Kappa Statistics.

From the Table No 6.1 and Figure No. 6.3 show the performance of Kappa Statistic on education dataset. We can see clearly that at 40% training size the Kappa statistic is -0.0187and when move from 40% to 50% the Kappa Statistic also increase. But when it goes to 60% the Kappa Statistic is decreasing and the value is 0.0237.At 70% and 80% the value of Kappa Statistic is regularly decreasing. Overall the value of Kappa Statistic is increase -0.0187to 0.0273.

C. Mean Absolute Error: The measure of the Mean Absolute Error of the mushroom dataset for BayesNet classifier techniques is shown below with graph according to the Table No6.1.

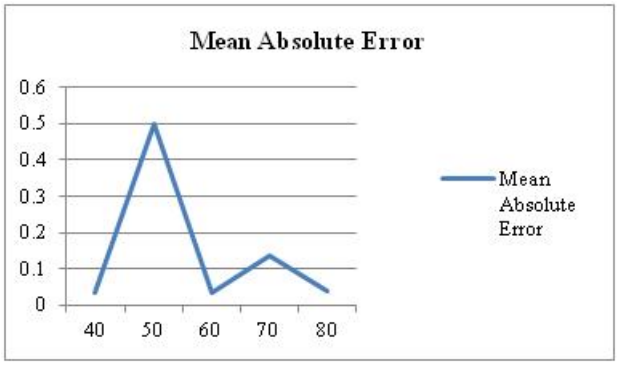


Figure No. 6.4 Show the Mean absolute Error

From the Table No.6.1 and Figure No.6.4 show the performance of the Mean Absolute Error on education dataset. From the Figure No.6.4 the Mean Absolute Error goes to highest to the lowest point. At the 40% training size the Mean Absolute Error is 0.0344; but at 50% training size Mean Absolute Error bit increase and at, 60% 70%and 80% Mean Absolute Error decrease. At 50% the Mean Absolute Error is the peak point and at the 40% the Mean Absolute Error is the lowest point.

7. NAÏVE BAYES:-

Naives is a simple form of Bayes net which is represented DAG with one parent and many children. Its use with a strong assumption of independence among child nodes in the context of their parent. [4] . Table no.7.1 shows the resultant measure the performance of the Naïve Bayes classifier techniques on the education dataset which have the total no. of instances is 624 and attributes are 16. We calculate the performance on the different training size (%) with the Naive Bayes techniques. Table No.7.1 show the training size in %, total no. of distinct Instances, correctly classified distinct instances, incorrectly classified distinct instances, distinct Mean Absolute Error and distinct Kappa Statistic.

Training Size (%)	Total no. of Instances (624)	Correctly Classified Instances % (value)	Incorrectly Classified Instances % (value)	Mean Absolute Error	Kappa Statistic
40	176	65.3409 %	34.6591 %	0.0312	0.0993
50	147	63.2653 %	36.7347 %	0.0315	0.0903
60	118	51.6949 %	51.6949 %	0.0324	0.0469
70	82	39.0244 %	60.9756 %	0.0332	0.0061
80	54	31.4815 %	68.5185 %	0.0349	-0.0544

Table No.7.1 Simulation result of algorithm Naive Bayes (Training)

A. *Accuracy*: The measure of the Accuracy of the education dataset for Naive Bayes classifier technique is shown below with graph according to the Table No.7.1 Based on the above Table No.7.1 and Figure No. 7.1, we can clearly see that the highest accuracy is 65.3409 and the lowest is 31.4815 %when the training set is 40% and 80% respectively in the Naïvebayes classifier technique. The other training sets yields an average accuracy of around 50%. we can clearly see that when the no. of instances are increasing the accuracy is increased. From the Figure No.7.1 the accuracy line has gone increase from 31% to65% approximately.

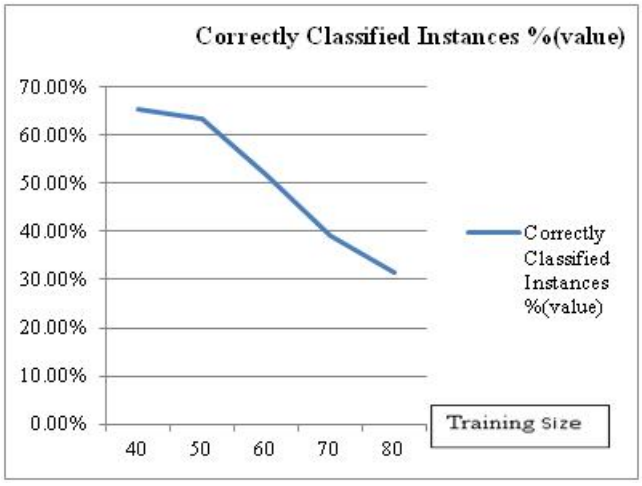


FIG 7.1 shows the performance of Accuracy.

B. *Kappa Statistic*: The measure of the Kappa Statistic of the mushroom dataset for BayesNet classifier techniques is shown below with graph according to the Table No.7.1. The Table No.7.1 and Figure No.7.2 show the performance of the Kappa Statistic applied the Naive Bayes classifier technique on the education dataset with the different size of the training set. From the Figure No. 7.2 we can see that at 40% training size the Kappa Statistic is 0.0993and when the training size is 80% the Kappa Statistic is -

0.0544.When dataset is small the Kappa Statistic is high and when the dataset is large the Kappa Statistic is low.

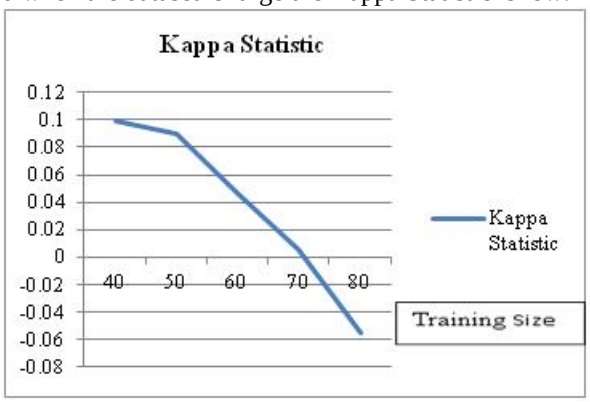


Figure No. 7.2 Show the Kappa Statistics.

C. *Mean Absolute Error*: The measure of the Mean Absolute Error of the education dataset for Naive Bayes classifier techniques is shown below with graph according to the Table No.7.1 The Table No7.1 and Figure No7.3 show the performance of the Mean Absolute Error applied the Naive Bayes classifier technique on the education dataset with the different size of the training set. From the Figure no.7.3 the mean absolute error goes from minimum to the maximum point. When the dataset training size is 40%, then error rate is low and when the dataset training size is 80% then error rate is high. From this figure we can see clearly that when the dataset is small the error rate is low and when dataset is large then error rate is high. It is same as the dataset size. From the figure No. 7.3 it is observed that the mean absolute error is increasing as well as the training size is increased.

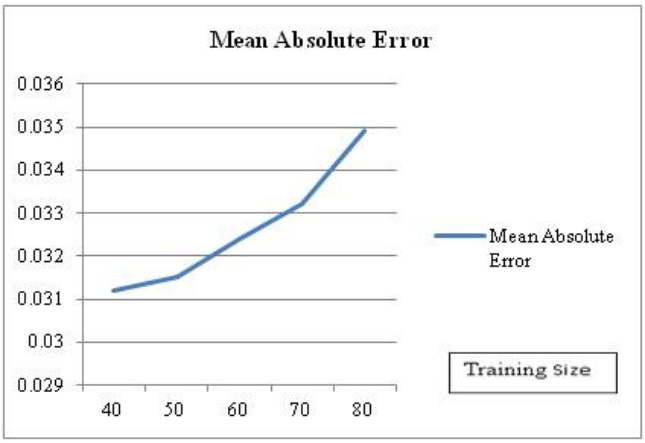


Figure No. 7.3 Show the Mean absolute Error.

8. DECISION TREES (DT'S):-

A decision tree is a tree where each non-terminal node represents a test or decision on the considered data item. Selection of a certain branch depends upon the outcome of the test [5]. . Table no.8.1 shows the resultant measure the performance of the classifier techniques on the education dataset which have the total no. of instances is 624 and attributes are 16. We calculate the performance on the different training size (%) with the Decision Trees (DT's) techniques. Table No.8.1 show the training size in %, total no. of distinct Instances, correctly classified distinct instances, incorrectly classified distinct instances, distinct Mean Absolute Error and distinct Kappa Statistic.

A. *Accuracy*: The measure of the Accuracy of the education dataset for Decision Trees (DT's) classifier technique is shown below with graph according to the Table No. 8.1From the below Table No.8.1 shows the training size (%), total no. instances, correctly classified instances and Kappa Statistic and Figure No.8.1 show the performance of Accuracy on education dataset. From the below Figure No.8.1. we can clearly see that the highest accuracy is 80.5085 %and lowest is 77.7778 %when the training size is 60% and 80% respectively. The accuracy sometimes increases and sometimes decreases on the different types of training size.

Training Size (%)	Total no. of Instances (624)	Correctly Classified Instances %(value)	Incorrectly Classified Instances %(value)	Mean Absolute Error	Kappa Statistic
40	176	79.5455 %	20.4545 %	0.014	0
50	147	78.9116 %%	21.0884 %	0.0141	0
60	118	80.5085 %	19.4915 %	0.0141	0
70	82	80.4878 %	19.5122 %	0.014	0
80	54	77.7778 %	22.2222 %	0.0146	0

Table No.8.1 Simulation result of algorithm decision tree (Training)

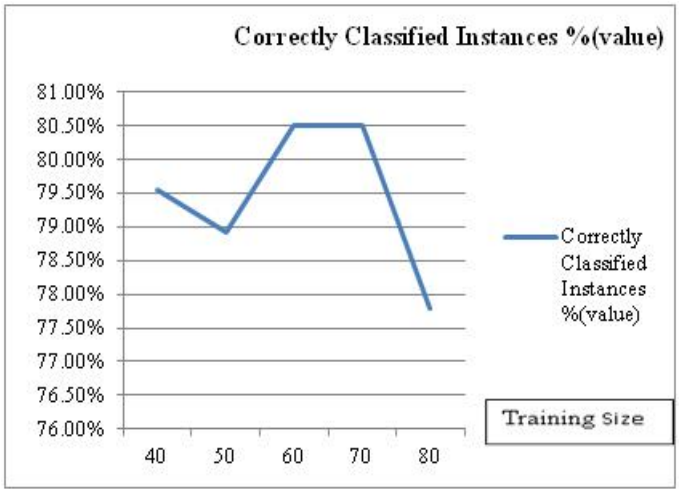


FIG 8.1 show the performance of Accuracy.

B. *Mean Absolute Error*: The measure of the Mean Absolute Error of the education dataset for Decision trees classifier techniques is shown below with graph according to the Table No.8.1 The Table No8.1 and Figure No8.2 show the performance of the Mean Absolute Error applied the Decision tree classifier technique on the educaton dataset with the different size of training set. From the Figure no.8.2 the mean absolute error is go from minimum to maximum point. When the dataset training size is 40% then error rate is low and when the dataset training size is 80% then error rate is high. From this figure we can see clear that when the dataset is small the error rate is low and when dataset is large then error rate is high. It is same as the dataset size. From the figure No. 7.3 it is observed that the mean absolute error is increase as well as the training size is increase.

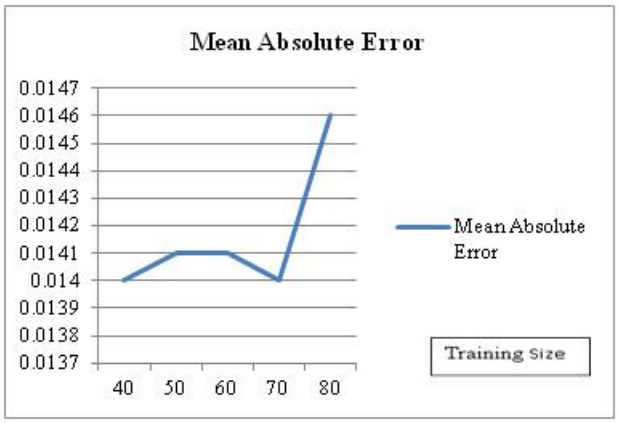


Figure No. 7.3 Show the Mean absolute Error.

09.COMPARISON OF DIFFERENT CLASSIFICATION TECHNIQUES

Here we compare different-2 techniques of classification in term of there accuracy , kappa statistics and mean square error.

A. *Comparison for accuracy:-*

Now we are describe the experimental results which is obtained from the various classification techniques and comparison with each other. The best techniques identified from each classifier then compared with other classifiers to discover what classifier is best to be used for classification of education dataset. We used here these techniques for the comparison on the education dataset and find the best techniques.

S. No.	Training Size (%)	BayesNet (%)	NaiveBayes (%)	Decision tree (%)
1	40	28.9773 %	65.3409 %	79.5455 %
2	50	23.1293 %	63.2653 %	78.9116 %%
3	60	16.1017 %	51.6949 %	80.5085 %
4	70	17.0732 %	39.0244 %	80.4878 %
5	80	9.2593 %	31.4815 %	77.7778 %

Table No 9.1 Comparative result of classification techniques

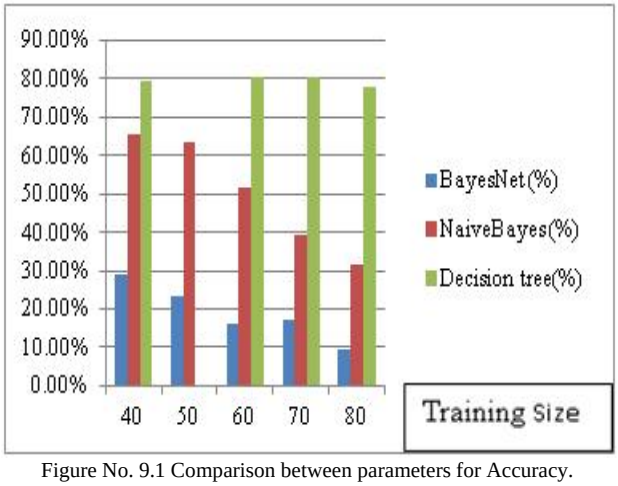


Figure No. 9.1 Comparison between parameters for Accuracy.

Based on the above Figure No.9.1 and Table no.9.1 we can clearly see that the highest accuracy is 28.9773 %and lowest accuracy is 9.2593 %in the BayesNet classifiers. And the highest accuracy is 65.3409 %and lowest accuracy is 31.4815 %%in the NaiveBayes classifiers and the highest accuracy is 80.5085 % and lowest accuracy is 77.7778 %in the Decision tree classifiers. From the above graph we can clearly see that the accuracy rate of Decision tree classifier is the best among these three classifier techniques.

B. *Comparison for Mean absolute Error:-*

S. No.	Training Size (%)	BayesNet (%)	NaiveBayes (%)	Decision tree(%)
1	40	0.0344	0.0312	0.0146
2	50	0.4986	0.0315	0.0141
3	60	0.0358	0.0324	0.0141
4	70	0.1371	0.0332	0.014
5	80	0.0374	0.0349	0.0146

Table No9.2 Comparative result of classification techniques

Based on the below Figure No. 9.2 and Table no. 9.2 we can clearly see that the highest Mean absolute error is 0.4986and lowest Mean absolute error is 0.0344% in the BayesNet classifiers. And the highest Mean absolute error is 0.0349and lowest Mean absolute error is 0.z0312in the NaiveBayes classifiers and the highest Mean absolute error is 0.0146% and lowest Mean absolute error is 0.014% in the Decision tree classifiers. From the above graph we can clearly see that the Mean absolute error rate of Bayes net classifier is the best among these three classifier techniques.

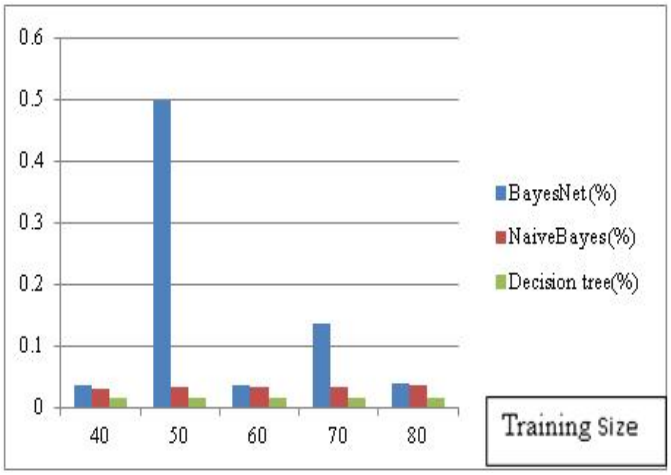


Figure No. 9.2 Comparison between parameters for Mean Absolute Error

C. *Comparison for Kappa Statistic:-*

S. No.	Training Size (%)	BayesNet (%)	NaiveBayes (%)	Decision tree(%)
1	40	-0.0187	0.0993	0
2	50	0.0273	0.0903	0
3	60	0.0237	0.0469	0
4	70	0.0103	0.0061	0
5	80	-0.0023	-0.0544	0

Table No9.3.Comparative result of classification techniques

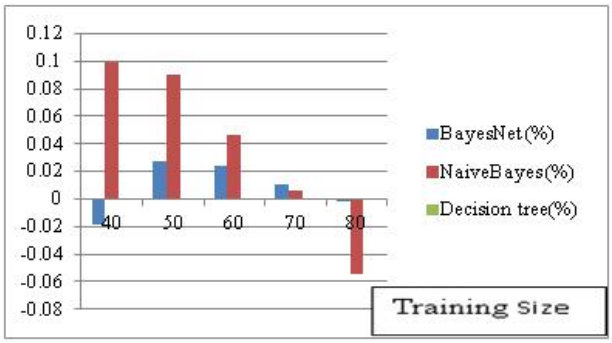


Figure No. 9.3Comparison between parameters for Kappa Statistic

Based on the above Figure No9.3 and Table no. 9.3we can clearly see that the highest Kappa Statistic is 0.0273and lowest Kappa Statistic is -0.0023in the BayesNet classifiers. And the highest Kappa Statistic is 0.0993and lowest Kappa Statistic is -0.0544in the NaiveBayes classifiers and the Kappa Statistic is zero in the Decision tree classifiers for all training size. From the above graph we can clearly see that the Kappa Statistic rate of Naïve net classifier is the best among these three classifier techniques.

CONCLUSION:-

After analyzing the results of testing the algorithms we can say that every techniques perform best result according to their parameters means if we take accuracy than decision tree is best but if we use mean absolute error than bayes net is better that other algorithms but if we take kappa statistics than naïve net perform better result.. Bayes network classifier has the potential to significantly improve the conventional classification methods for use in general education field.

REFERENCES

[1]. Dr. Mohd Maqsood Ali, "ROLE OF DATA MINING IN EDUCATION SECTOR"International Journal of Computer Science and Mobile Computing Vol.2 Issue. 4, April- 2013, pg. 374-383.

[2]. .Sapna Jain, M Afshar Aalam and M N Doja, "K-means clustering using weka interface", Proceedings of the 4th National Conference; INDIACom-2010.

[3]. . Thair Nu Phyu"Survey of Classification Techniques in Data Mining" Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I 2009, March 18 - 20, 2009, Hong Kong.

[4]. S. B. Kotsiantis · I. D. Zaharakis · P. E. Pintelas" Machine learning: a review of classification and combining techniques" Published online: 10 November 2007 © Springer Science+Business Media B.V. 2007

[5]. A.Shameem Fathima[1] ,D.Manimegalai[2] and Nisar Hundewale "A Review of Data Mining Classification Techniques Applied for Diagnosis and Prognosis of the Arbovirus-Dengue" IJCSI International Journal of sComputer Science Issues, Vol. 8, Issue 6, No 3, November 2011.

[6]. Abdullah H. Wahbeh, Qasem A. Al-Radaideh, Mohammed N. Al-Kabi, and Emad M. Al-Shawakfa" A Comparison Study between Data Mining Tools over some Classification Methods" *(IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence*

[7]. Bharat Chaudhari1, Manan Parikh2" A Comparative Study of clustering algorithms Using weka tools" *International Journal of Application or Innovation in Engineering & Management (IJAIEM)* Web Site: www.ijaiem.org Email:editor@ijaiem.org,editorijaiem@gmail.comVolume 1, Issue 2, October 2012.