International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 4, Issue 3, May-June 2015 **ISSN 2278-6856**

A Comparative Performance Analysis of Classifications Methods in Data mining On Medical data set using the weka tool

[1]Geeta , Mr. Trilok Gaba[2]

[1] Student of Masters of Technology, B.I.Ts (M.D. University) Rohtak, Haryana, India

[2]Assistant Professor B.I.Ts College (M.D. University) Rohtak, Haryana, India

Abstracts

We can se e our surrounding many diseases are spread due to many reasons and every diseases have some sytsemous and treatments .due to theses dieses many people are affected Thus we need data mining to store the data . 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 holds great potential for the medical industry to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs. Some experts believe the opportunities to improve care and reduce costs concurrently could apply to as much as 30% of overall healthcare spending. . Classification is one of them technique of data mining, which used to predefine classification data. 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

Databases with patient health information have been used for a long time and have not been considered to create any problems. Databases have been used in various ways such as for quality control and research and health authorities have for some time used databases for citizen health information. It is important for example for the authorities to have information on infectious diseases that may cause much distress in society. Each database has a delimited and defined purpose even though in many cases they contain a large amount of information. A formal definition of Knowledge discovery in databases is given as follows: "Data mining is the non trivial extraction of implicit previously unknown and potentially useful information about data" [7]. The availability of integrated information via the huge patient repositories, there is a shift in the perception of clinicians, patients and payers from qualitative visualization of clinical data by demanding a more quantitative assessment of information with the supporting of all clinical and imaging data. For instance it might now be possible for the physicians to compare diagnostic information of various patients with identical conditions. Likewise, physicians can also confirm their findings with the conformity of other physicians dealing with an identical case from all over the world [8]. Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. The automation of the same would be highly beneficial.

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]. In data classification we use learning and classification. Classification is classified into different
models, these are followed:-
Types of classification models:-
o Classification by decision tree induction
o Bayesian Classification
o Neural Networks
o Support Vector Machines (SVM)
o Classification Based on Associations.

3. WEKA TOOL

Weka is a landmark tool 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[6].

Figure3. 1View of weka tool.

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 medical data set.

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. In this data having 367instances and 23 attributes.

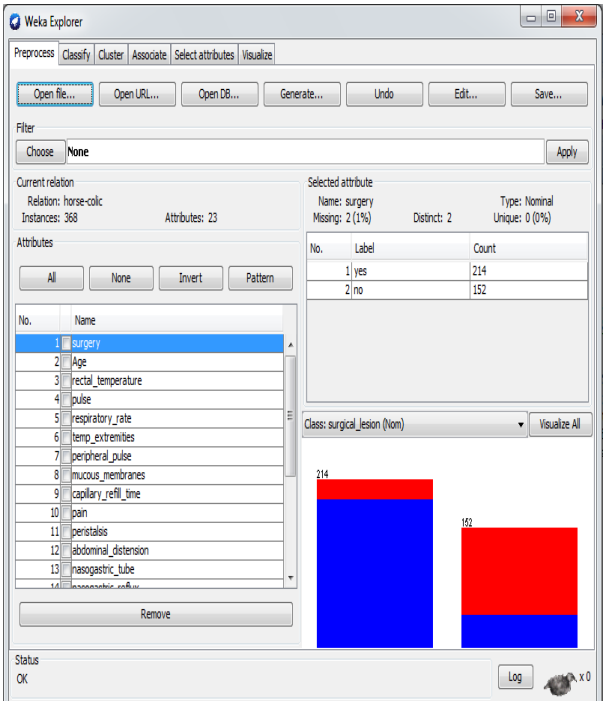
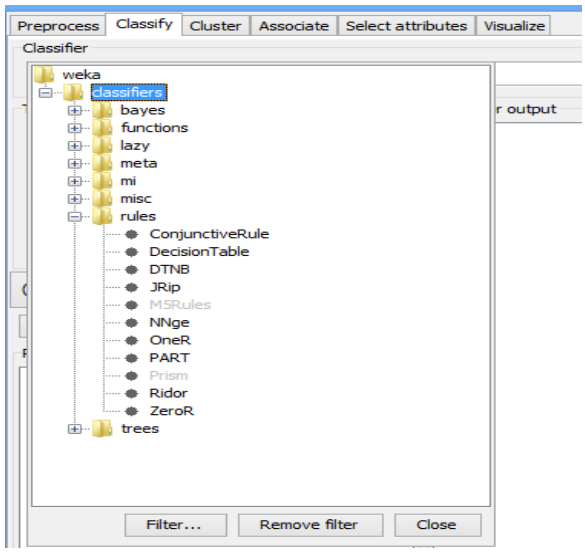


Figure 5.1: load data set in to the weka

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].

A. Accuracy: The measure of the Accuracy of the medical dataset for BayesNet classifier technique is shown below with graph according to the Table No. 6.1. 6.1From the below Table No.6.1 shows the name of data set, total no. instances, correctly classified instances and Kappa Statistic and Figure No.6.2 show the performance of Accuracy on medical dataset Figure No.6.2 show the performance of Accuracy on Medical dataset

Table 6. 1. Show the result of Bayes net

Dataset Name	Correctly Classified Instances % (value)	Incorrectly Classified Instances % (value	Kappa Statistic	Mean Absolute Error
colic	81.25 %	18.75 %	0.5994	0.2168
diabetes	74.349 %	25.651 %	0.429	0.2987
hypothyroid	98.5949 %	1.4051 %	0.9028	0.011

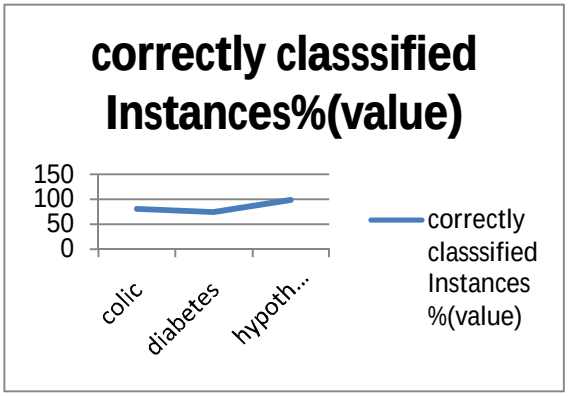


FIG 6.2 show the performance of Accuracy.

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

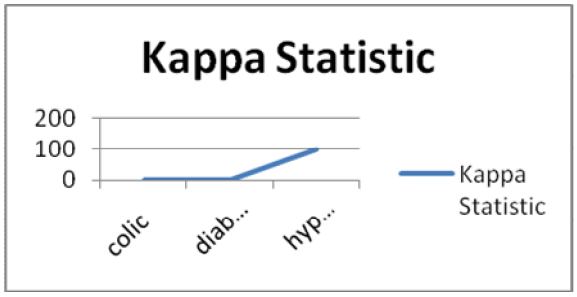


Figure No. 6.3 Show the Kappa Statistics.

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

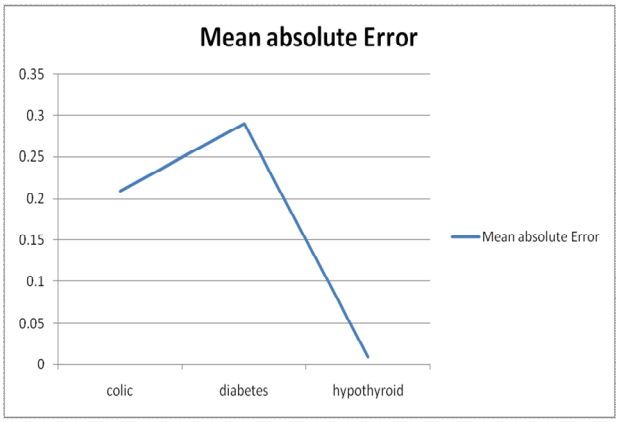


Figure No. 6.4 Show the Mean absolute Error

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 medical dataset.

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

Dataset Name	Correctly Classified Instances % (value)	Incorrectly Classified Instances % (value	Kappa Statistic	Mean Absolute Error
colic	77.9891 %	22.0109 %	0.53	0.23
diabetes	76.3021 %	23.6979 %	0.46	0.28
hypothyroid	95.281 %	4.719 %	0.60	0.035

A. Accuracy: The measure of the Accuracy of the Medical 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. 7.1Based on the above Table No.7.1 and Figure No. 7.1, we can clearly see that the highest accuracy is 95% and the lowest is76%.

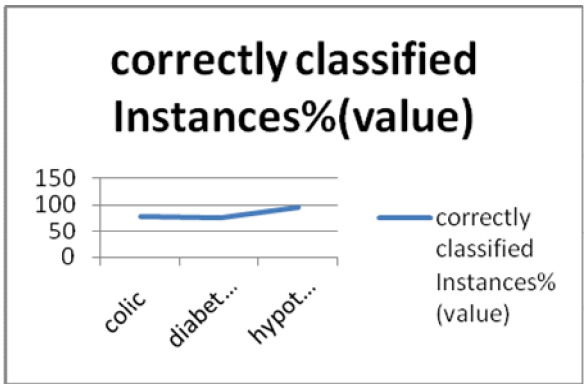


FIG 7.1 shows the performance of Accuracy.

B. Kappa Statistic: The measure of the Kappa Statistic of themedicaldataset 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 medical dataset with the different size of the training set.

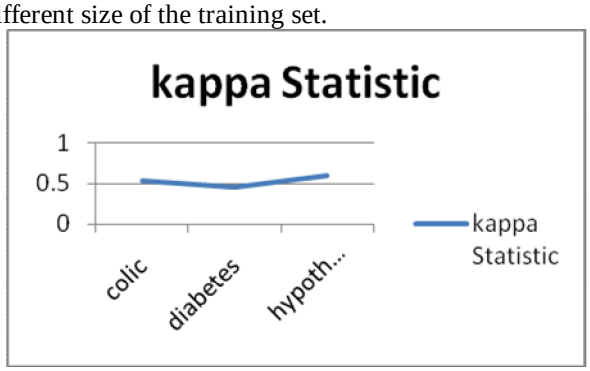


Figure No. 7.2 Show the Kappa Statistics

C. Mean Absolute Error: The measure of the Mean Absolute Error of the medical 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 medical 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.

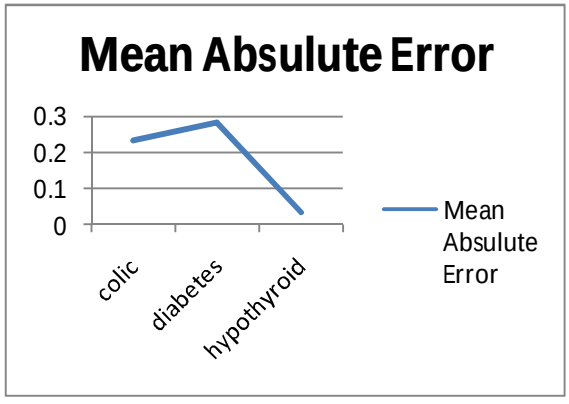


Figure No. 7.3 Show the Mean absolute Error.

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 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.

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

Dataset Name	Correctly Classified Instances % (value)	Incorrectly Classified Instances % (value	Kappa Statistic	Mean Absolute Error
colic	85.3261 %	14.6739 %	0.6732	0.24
diabetes	73.8281 %	26.1719 %	0.41	0.31
hypothyroid	99.5758 %	0.4242 %	0.97	0.03

A. Accuracy: The measure of the Accuracy of the medical 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 medical dataset.

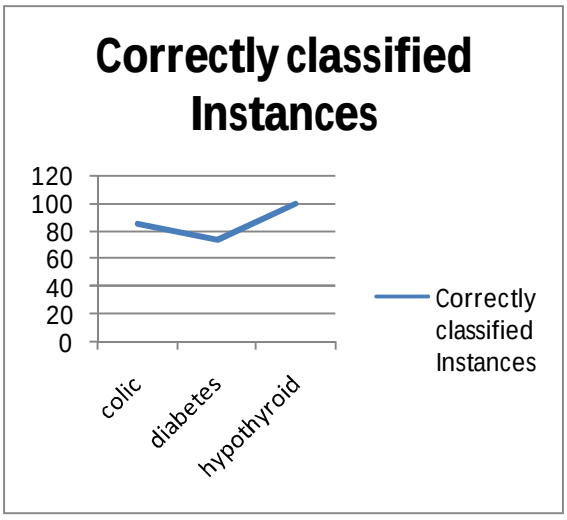


FIG 8.1 show the performance of Accuracy

B. Mean Absolute Error: The measure of the Mean Absolute Error of medical 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 medical dataset with the different size of training set. From the Figure no.8.2 the mean absolute error is go from maximum to minimum

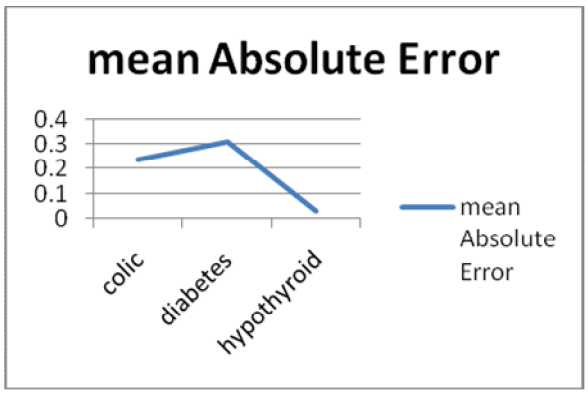


Figure No. 8.2 Show the Mean absolute Error.

9.COMPARISON OF DIFFERENT CLASSIFICATION TECHNIQUE

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 medical dataset. We used here these techniques for the comparison on the medical dataset and find the best techniques.

Table No 9.1 Comparative result of classification techniques

Sr.no.	Data set name	BayesNet %	NaiveBayes (%)	Decision tree(%)
1	colic	81.25 %	77.9891 %	85.3261 %
2	diabetes	74.349 %	76.3021 %	73.8281 %
3	hypothyroid	98.5949 %	95.281 %	99.5758 %

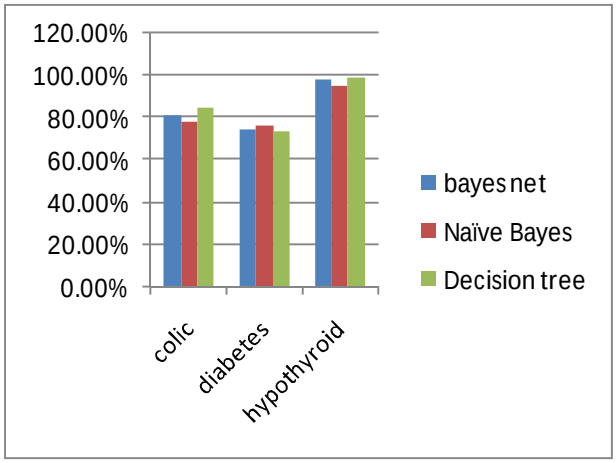


Figure No. 9.1 Comparison between parameters for Accuracy.

Based on the above Figure No.9.1 and Table no.9.1 we canclearly see that the highest accuracy is 99.5 %and lowest73 % of decision tree.

B. Comparison for Mean absolute Error

Table No9.2 Comparative result of classification techniques

Sr.no.	Data set name	BayesNet	NaiveBayes	Decision tree
1	Colic	0.2168	0.23	0.24
2	Diabetes	0.2987	0.28	0.31
3	hypothyroid	0.011	0.035	0.03

Based on the below Figure No. 9.2 and Table no. 9.2 we can clearly see that the highest Mean absolute error is 0.24986and lowest Mean absolute error is 0.0344%.

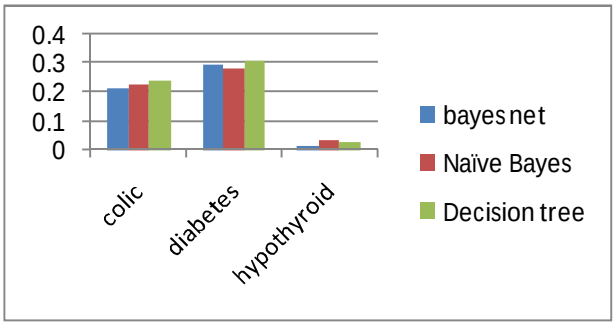


Figure No. 9.2 Comparison between parameters for Mean Absolute Error

C. Comparison for Kappa Statistic

Table No9.3. Comparative result of classification techniques

Sr.no.	Data set name	BayesNet	NaiveBayes	Decision tree
1	colic	0.5994	0.53	0.6732
2	diabetes	0.429	0.46	0.41
3	hypothyroid	0.9028	0.60	0.97

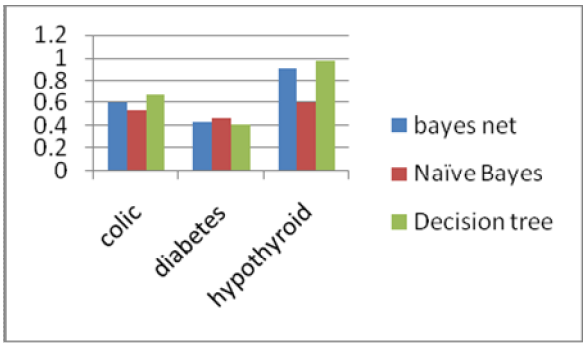


Fig.No9.3. Comparative result of classification techniques

Based on the above Figure No9.3 and Table no. 9.3we can clearly see that the highest Kappa Statistic is 1 lowest Kappa Statistic is 0.41.

10.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 can be to find the to significantly potential to improve the conventional classification methods for use in general medical 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.Manimegalai2 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]. Suman #1, Mrs.Pooja Mittal*2" A Comparative Performance Analysis of Classification Algorithms Using Weka Tool Of Data Mining Techniques" Suman et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3448-3453

[7]. Frawley and Piatetsky-Shapiro, 1996. Knowledge Discovery in Databases:An Overview. The AAAI/MIT Press, Menlo Park, C.A.

[8]. Miller, A., B. Blott and T. Hames, 1992. Review of neural network applications in medical imaging and signal processing. Med. Biol. Engg. Comp., 30: 449-464