International Journal For Technological Research In Engineering
Volume 5, Issue 9, May-2018 *ISSN (Online): 2347 - 4718*

PERCEPTIVE PREDICTION OF MACHINE LEARNING OVER BIG DATA

V.Malleeswari[1], Mrs.M.P.Sujatha[2]
[1]PG Student, Department of Computer Science,
[2]Associate Professor, Sri Ramanujar Engineering College, Tamil Nadu, India

Abstract: In the health care sector, large data collected every day and stored in databases to perform medical research and experiments. The challenge in health care sector is that for the same patient different doctors may diagnose in a different way. To overcome the challenges in healthcare sectors, automated classification tools are used to diagnose the disease accurately and efficiently and it may also reduce the time and burden on doctors. In this paper we analyze and compare the behaviour of five different classification algorithms namely Naive Bayes classifier, Decision tree algorithm, AdaBoost algorithm, K means clustering and Support Vector Machines on the given medical data sets. The datasets used for this work is an open source gathered from uci machine learning repository. The algorithms performance are measured based on eight criteria: Sensitivity, specificity, false positive rate, positive predictor value, negative predictor value, false discovery rate, accuracy and error rate.

Index Terms: AdaBoost algorithm, Naive bayes classifier, Support Vector Machine, Decision Tree, K means clustering.

I. INTRODUCTION

The vast amounts of data generated by healthcare industry are too complex to analyze by means of existing methods. When data mining is applied to the medical sector useful information can get extracted easily for further knowledge discoveries. With the help of data mining, information can be retrieved easily even for large datasets. Classification process is a tree based structure comes under data mining concept. Because of using various classification algorithms in data mining, disease can be predicted easily. In this paper the performance of five classification algorithms namely Naive Bayes Classifier (NBC), Support Vector Machines(SVM), Decision Tree, K means clustering AdaBoost are analyzed. The datasets considered for the project was pimaindians diabetes, breast cancer Wisconsin (original), Bupa liver disorder, statlog (heart) dataset. The datasets used for this project are open source gathered uci machine learning repository. The performance of algorithms are measured based on eight criteria namely sensitivity, specificity, false positiverate, positive predictor value, negative predictor value, false discovery rate, accuracy and error rate.

II. RELATED WORK

In the works of Menaka et al., [1] paper keywords are extracted from documents Using TF-IDF and WordNet. TF-IDF algorithms are used to select the candidate words and WordNet is a lexical database of English which is used to find similarity among the candidate words. The highest similarity in the words is taken as keywords. The selected classification algorithms used for their paper are Naïve Bayes, Decision tree and K-Nearest Neighbour (KNN) algorithms and its performance are analyzed. Based on the performance criteria it is found that decision tree algorithm gives the better accuracy for text classification among other algorithms. Day by day patient with liver disease increasing extraordinarily due to the excessive alcohol, inhale of harmful gases, intake of contaminated food, pickles and drugs. Bendi Venkata Ramana et al., [2] paper evaluates the selected classification algorithms for the liver patient datasets. The selected classification algorithms used for their project to check the liver disease in patient are Naive Bayes classifier, C4.5, Back propagation Neural Network algorithm, and Support Vector Machines. These algorithms are measured based on sensitivity, specificity, accuracy and precision. Gupta et al., [3] have used four classification algorithms namely J48, Random Forest (RF), Reduce Error Pruning (REP) and Logistic Model Tree (LMT) to classify the “WEATHER NOMINAL” open source Data Set. Waikato Environment for Knowledge Analysis (WEKA) has been used in their paper for evaluating the performance of classification algorithms and they found that Random Forest algorithms classify the given data set better than the other algorithms. The performance of classifier algorithms is experimented for 5-fold cross validation test. In the works of Nilakshi et al., [4] Heart Disease Prediction System is developed using Neural Network and Genetic Algorithms.

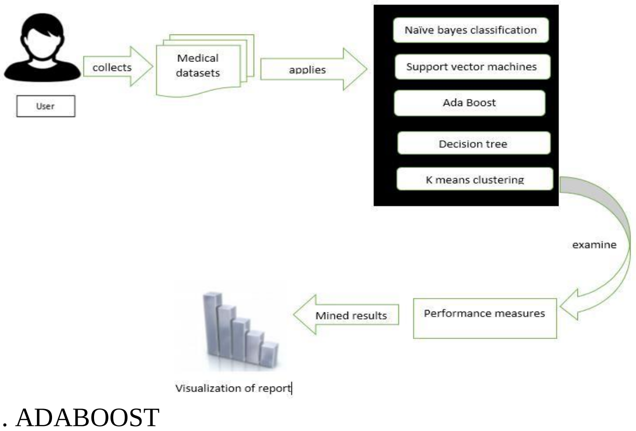
Prediction system calculates the number of hidden nodes for neural network which train the network with proper selection of neural network architecture and uses the global optimization of genetic algorithms for initialization of neural network. Based on experimented result it is found that genetic neural approach predicts the heart disease up to 98% accuracy. Mukesh kumari et al., [5] have used Bayesian Network classifier to predict the persons are having chance for diabetes. The dataset used was collected from a hospital, which tells about the diabetes information about the particular patient. Weka tool is used for the experiment and analysis. Bayesian network classifier algorithms is applied on the dataset of persons collected from hospital. Results have been obtained. Prediction of dermatological disease is very difficult. Because of the number of diseases presentation. Manjusha et al., [6] have proposed a system which allows obtaining data patterns with the help of Naive Bayesian theorem. The data experimented for this project was gathered from tertiary health care centers which survey the people from various areas of Kottayam and Alappuzha, Kerala, India. Naive Bayesian algorithms betray the chances of different dermatological disease and also finds out the percentage of occurrence of each disease. Vijayarani et al., [7] have used classification algorithms to check the presence of liver diseases in patients. Naive Bayes and support vector machine (SVM) are the two classification algorithms used in their paper to check for the occurrence of liver diseases. Accuracy and execution time are the two performance factors used in their paper to measure the effectiveness of algorithms. From the evaluated results it is found that the SVM is a better classifier when compared to Naive Bayes for predicting the liver diseases.

III. METHODOLOGY

Figure 1 shows the comparison of five different classification algorithms on the given medical datasets. Comparisons are done based on the performance factors such as sensitivity, specificity, false positive rate, positive predictor value, negative predictor value, false discovery rate, accuracy and error rate. Based on performance result, system will predict the best algorithm to classify the given data.

1. Decision Tree

Decision tree most commonly used in data mining. Decision tree can be applied to enormous amount of data. Decision tree is a flow chart like structure that includes a root node, branches, and leaf nodes. Each root node denotes a test on an attribute and it located at the top of the tree. Each branch denotes the outcome of a test and it located in the middle of the tree. Each leaf node holds a class label and it is located at the end. Pruning is the process performed in decision tree in order to remove anomalies in the training data due to noise or outliers. Two approaches involved in tree pruning. Pre pruning is a process stopping its construction early whereas post pruning removes a sub tree from fully grown tree.



2. ADABOOST

AdaBoost name comes from "Adaptive Boosting", and it is one of the popular machine learning algorithms. It can be used in association with many other types of learning algorithms to improve their ability. The output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final outcome of the boosted classifier. AdaBoost is an algorithm for constructing strong classifier from the given weak learners. AdaBoost is adaptive because of subsequent weak learners are tweaked in favour of those instances misclassified by previous classifiers. AdaBoost is sensory to noisy data and outliers. In some problems it can be less vulnerable to the overfitting problem than other learning algorithms.

3. Support Vector Machine

Support vector machine are supervised learning models that constructs a hyper plane in a high-dimensional space, which can be used to analyze data for classification, regression analysis and outliers. A good separation is achieved by the hyper plane that separates the objects into respective groups. The advantages of support vector machines are:

- When number of dimensions is greater than the number of samples it is very effective.
- Because of using subset of training points in the decision function (called support vectors), it is said to be memory efficient.

The disadvantages of support vector machines are:

- When number of features is greater than the number of samples, the method is expected to give poor performances.
- Instead of using probability estimates, it uses an expensive five-fold cross-validation.

4. K means clustering algorithm

k-means clustering is a method of vector quantization, came originally from signal processing, it is a popular algorithm for doing cluster analysis in mining. Clustering is the process of partitioning a group of data points into a small number of clusters. K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters fixed a priori. The following are the steps involved in k means clustering

- Cluster the data in to k groups where k is predefined.
- Select k points at random as cluster centers
- Assign objects to the closest cluster center according to the Euclidean distance function.
- Calculate centroid or mean of all objects in each cluster
- Repeat steps two, three and four until the same points are assigned to each cluster in consecutive round.

5. Naive Bayes classification algorithm

In machine learning technique naive Bayes classifiers is a family of simple probabilistic classifiers based on applying Bayes theorem with strong independence assumptions between the features. Naive Bayes is a conditional probability model given a problem instance to be classified, Represented by a vector x=(x_1, …,x_n) representing some n features(independent variables), it assigns to these instance probabilities

$$p(C_k|x_1...,x_n) \qquad (1)$$

For each of K possible outcomes or class

The above formula is not convenient when number of features is large. To make it amenable using Bayes theorem the above formula can be decomposed as

$$P\left(C_k/X\right) = \frac{P(C_k)\,P\left(X/C_k\right)}{P(X)} \qquad (2)$$

To make it understandable the above formula can be written as

$$posterior = \frac{prior \times likelihood}{evidence} \qquad (3)$$

The following are the application of Naive Bayes algorithms

- Real time Prediction**:** Naive Bayes can make prediction in real time.
- Multi class Prediction: It can even suitable for predict the probability of multiple classes of target variable.
- Text classification/ Spam Filtering/ Sentiment Analysis: Naive Bayes classifiers mostly used in text classification have higher success rate as compared to other algorithms. As a result, it is widely used in Spam filtering (identify spam e-mail) and Sentiment Analysis (in social media analysis, to identify positive and negative customer sentiments)

IV. DATA SETS DESCRIPTION

The detailed description of four datasets are given below

1. Pimaindians diabetes Dataset The data was collected from a population of women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona, who were tested for diabetes according to World Health Organization criteria. The data were collected by the US National Institute of Diabetes and Digestive and Kidney Diseases. The detailed description of pimaindians diabetes dataset available in uci machine learning repository [8]. The data consists of 768 observations on 9 variables. The attributes for the pimaindians diabetes dataset are given below

1) Number of times pregnant
2) Plasma glucose concentration a 2 hours in and oral glucose tolerance test
3) Diastolic blood pressure (mm Hg)
4) Triceps skin fold thickness (mm)
5) 2-hour serum insulin (mu U/ml)
6) Body mass index (weight in kg/ (height in m) ^2)
7) Diabetes pedigree function
8) Age (years)
9) Class variable (0 or 1)

2.Wisconsin Breast Cancer Dataset

Breast Cancer Wisconsin (Original) Data Set was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg.The detailed description of this dataset available in uci machine learning site [9]. There are 699 observations in the data and eleven attributes including class attribute. Class attribute has one of 2 possible classes: benign or malignant.

The attributes of breast cancer dataset are

1.Sample code number - Id number
2.Clump thickness - 1-10
3.Uniformity of cell size - 1-10
4.Uniformity of cell shape - 1-10
5.Marginal adhesion - 1-10
6.Single epithelial cell size- 1-10
7.Bland chromatin - 1-10
5.Noramal nucleoli - 1-10
9.Mitoses - 1-10
10.Class - 1 for benign and 2 for malignant

3. BUPA liver disorders Dataset

The data was gathered from BUPA Medical Research Ltd. The dataset contains 345 instances on 7 variables including class attribute. Four duplicate values were removed in the dataset by data cleaning. The first 5 variables are all blood tests which are thought to be sensitive to liver disorders that might arise from excessive alcohol consumption. Each line in the Bupa. Data file constitutes the record of a single male individual. The detailed description of this dataset available in uci machine learning [10].

The attributes of the liver disorders datasets are given below

1.mcv - mean corpuscular volume
2.alkphos - alkaline phosphatase
3.sgpt - alamine minotransferase
4.sgot - aspartate inotransferase
5.gammgt gamma - glutamyl transpeptidase
6.drinks - Number of alcoholic beverages drunk per day.

4. *Heart (statlog) dataset*

This dataset consists of 270 instances and 14 attributes including class attribute. The detailed description of this dataset available in uci machine learning repository [11]. The attribute information of Statlog (Heart) Data Set are given below.

Attribute Information:

1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholesterol in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina
10.oldpeak = ST depression induced by exercise relative to rest
4. the slope of the peak exercise ST segment
5. number of major vessels (0-3) colored by fluoroscopy
6. thal: 3 = normal; 6 = fixed defect; 7 = reversible defect
7. Class-Absence (1) or presence (2) of heart disease.

V. RESULTS AND DISCUSSION

Performance of selected classification algorithms were experimented with four datasets such as pimaindians

diabetes, breast cancer Wisconsin(original), statlog (heart data set) and Bupa liver disorders datasets. The four datasets are open sourced gathered from uci machine learning site. In each and every algorithm we have noticed the parameters sensitivity, specificity, false positive rate, positive predictor value, negative predictor value, false discovery rate, accuracy and error rate.
The parameters illustrated as follows

Sensitivity-It is used to measure the performance of true positive rate

$$sensitivity = \frac{TP}{TP + FN}$$

Specificity-It is used to measure the performance of true negative rate

$$specificty = \frac{TN}{TN + FN}$$

False positive rate- It is the ratio of individual who incorrectly received a positive test result

$$False\, positive\, rate = \frac{FP}{FP + TN}$$

Positive predictor value- If the test result is positive what is the probability that the patient actually has the disease

$$Positive\, predicted\, value = \frac{TP}{TP + FP}$$

Negative predictor value-If the test result is negative what is the probability that the patient does not have the disease

$$Negative\, predicted\, value = \frac{TN}{TN + FN}$$

False discovery rate-It is a way of conceptualizing the *rate* of type I errors in null hypothesis testing when conducting multiple comparisons

$$False\, discovery\, rate = \frac{FP}{FP + TP}$$

Accuracy-It is defined as the ratio of correctly classified instances to total number of instances

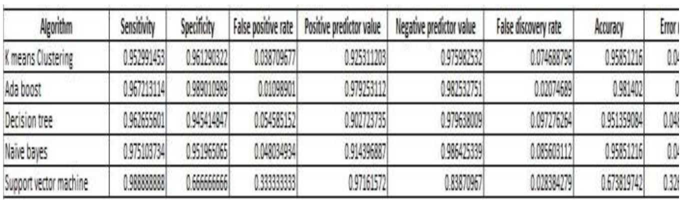
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Error rate-It is the number of bit errors per unit time. It can be calculated with the help of accuracy as shown below

$$Error\, rate = 1 - accuracy$$

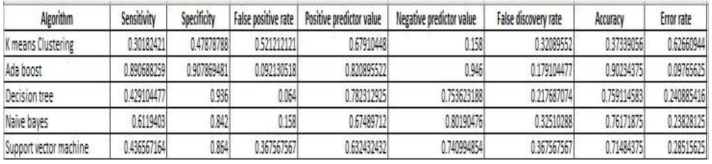
Big data analytics has helped health care improve by providing personalized medicine and prescriptive analytics, clinical risk intervention and predictive analytics, waste and care variability reduction, automated external and internal reporting of patient data, standardized medical terms and patient registries and fragmented point solution. Efficiently handling large volumes of medical imaging data and extracting potentially useful information and biomarkers.

Table 1: Performance measures of Pimaindians diabetes dataset



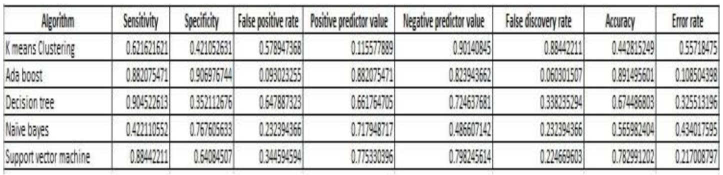
Algorithm	Sensitivity	Specificity	False positive rate	Positive predictor value	Negative predictor value	False discovery rate	Accuracy	Error r
K means Clustering	0.952991453	0.961290322	0.038709677	0.925311203	0.975982532	0.074688796	0.95851216	0.04
Ada boost	0.967213114	0.989010989	0.01098901	0.979253112	0.982532751	0.02074689	0.981402	0
Decision tree	0.962655601	0.945414847	0.054585152	0.902723735	0.979638009	0.097276264	0.951359084	0.04
Naïve bayes	0.975103734	0.951965065	0.048034934	0.914396887	0.986425339	0.085603112	0.95851216	0.04
Support vector machine	0.988888888	0.666666666	0.333333333	0.97161572	0.83870967	0.028384279	0.673819742	0.32

Table 2: Performance measures of breast cancer Wisconsin (original dataset)



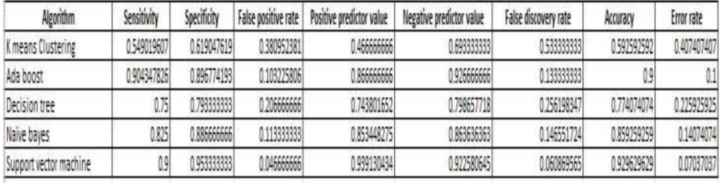
Algorithm	Sensitivity	Specificity	False positive rate	Positive predictor value	Negative predictor value	False discovery rate	Accuracy	Error rate
K means Clustering	0.30182421	0.47878788	0.521212121	0.67910448	0.158	0.32089552	0.37339056	0.62660944
Ada boost	0.890688259	0.907869481	0.092130518	0.820895522	0.946	0.179104477	0.90234375	0.09765625
Decision tree	0.429104477	0.936	0.064	0.782312925	0.753623188	0.217687074	0.759114583	0.240885416
Naïve bayes	0.6119403	0.842	0.158	0.67489712	0.80190476	0.32510288	0.76171875	0.23828125
Support vector machine	0.436567164	0.864	0.367567567	0.632432432	0.740994854	0.367567567	0.71484375	0.28515625

Table 3: Performance measures of Bupa liver disorder dataset



Algorithm	Sensitivity	Specificity	False positive rate	Positive predictor value	Negative predictor value	False discovery rate	Accuracy	Error rate
K means Clustering	0.621621621	0.421052631	0.578947368	0.115577889	0.90140845	0.88442211	0.442815249	0.55718475
Ada boost	0.882075471	0.906976744	0.093023255	0.882075471	0.823943662	0.060301507	0.891495601	0.108504398
Decision tree	0.904522613	0.352112676	0.647887323	0.661764705	0.724637681	0.338235294	0.674486803	0.325513196
Naïve bayes	0.422110552	0.767605633	0.232394366	0.717948717	0.486607142	0.232394366	0.565982404	0.434017595
Support vector machine	0.88442211	0.64084507	0.344594594	0.775330396	0.798245614	0.224669603	0.782991202	0.217008797

Table 4: Performance measures of statlog (heart)dataset



Algorithm	Sensitivity	Specificity	False positive rate	Positive predictor value	Negative predictor value	False discovery rate	Accuracy	Error rate
K means Clustering	0.549019607	0.619047619	0.380952381	0.466666666	0.693333333	0.533333333	0.592592592	0.407407407
Ada boost	0.904347826	0.896774193	0.103225806	0.866666666	0.926666666	0.133333333	0.9	0.1
Decision tree	0.75	0.793333333	0.206666666	0.743801652	0.798657718	0.256198347	0.774074074	0.225925925
Naïve bayes	0.825	0.886666666	0.113333333	0.853448275	0.863636363	0.146551724	0.859259259	0.14074074
Support vector machine	0.9	0.953333333	0.046666666	0.939130434	0.922580645	0.060869565	0.929629629	0.07037037

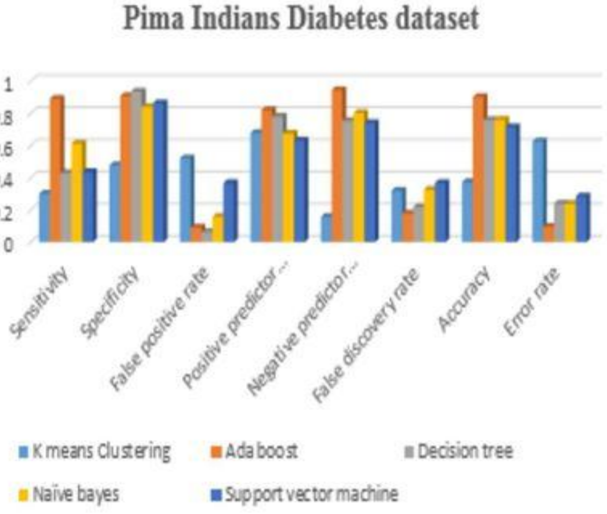


Figure 2: Visualization graph

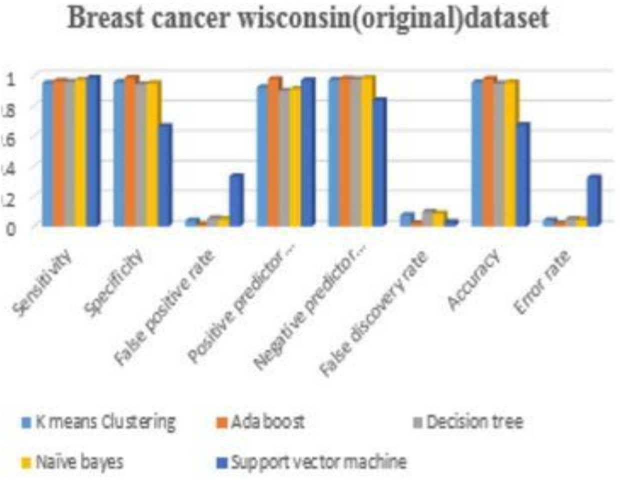


Figure 3: Visualization graph

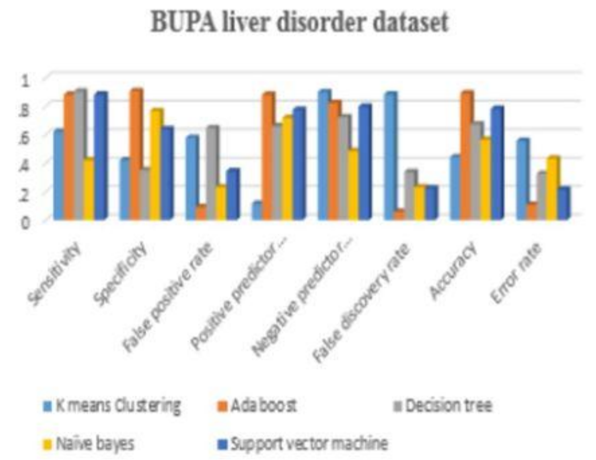


Figure 4: Visualization graph

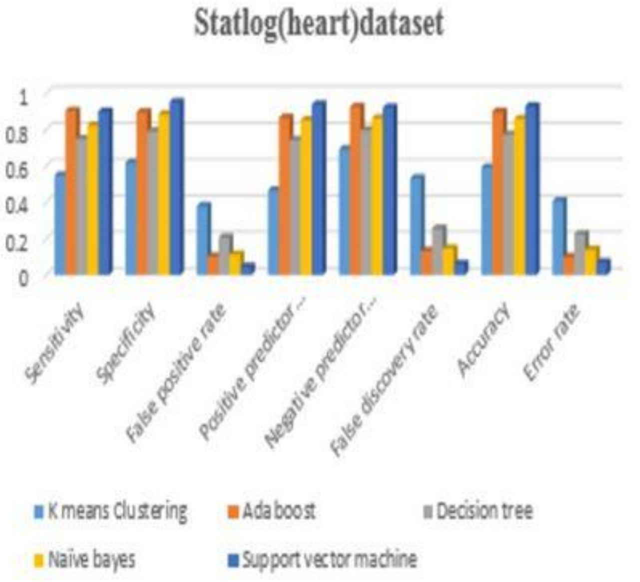


Figure 5: Visualization graph

In this paper, the results are analyzed with various classification algorithms on the given medical datasets. This work is implemented in R tool 3.2.2 version. Figure 2 represents the performance measures of classification algorithms on pimaindiansdiabetes dataset. An evaluated result shows the performance of adaboost is better than other algorithms. Figure 3 represents performance measures of classification algorithms on breast cancer Wisconsin(original) dataset. An evaluated result shows the performance of adaboost is better than other algorithms. Figure 4 represents performance measures of classification algorithms on Bupa liver disorder dataset. An evaluated result shows the performance of adaboost is better than other algorithms. Figure 5 represents performance measures of classification algorithms on statlog (heart) dataset. An evaluated result shows the performance of svm is better than other algorithms.

VI. CONCLUSION AND FUTURE WORK

In this paper we have used R tool for analysis and Performed comparison of selected classification algorithms. After the comparative analysis we examined that adaboost algorithms is more accurate and has less error rate. In our paper we also represented graphically and finally concluded that adaboost is best among the all other algorithms. Our interface also provides the user the choice of selecting suitable prediction algorithm. More similar studies on different data set for machine learning approach is needed to confirm the above finding.

REFERENCES

[1] Menaka.S, Radha.N,” Text Classification using Keyword extraction technique”, International Journal of Advanced Research in Computer Science and Engineering, Volume 3, issue12, pp.734-740, December 2013.

[2] Bendi Venkata Ramana, Surendra Prasad Babu and N.B Venkateswarlu,"A Critical Study of Selected Classification", International Journal of Database Management Systems, Volume 3, No.2, pp.101-114, May2011.

[3] D. LGupta, Malviya and Satyendrasingh,” Performance Analysis of Classification Tree Learning Algorithms”, International Journal of Computer Application, Volume 55, No.6, pp.39-44,2012.

[4] Nilakshi. P, Waghulde1, Nilima. P and Patil, "Genetic Neural Approach for Heart Disease Prediction”, International Journal of Advanced Computer Research, Volume 4, No.3, issue 16, pp.778-784, September-2014.

[5] Mukesh kumari, Dr. Rajan Vohra, and Anshul arora,” Prediction of Diabetes Using Bayesian Network”, International Journal of Computer Science and Information Technologies, Vol. 5 (4), pp. 5174-5178,2014.

[6] ManjushaK, Sankaranarayanan.k, Seena.P,” Prediction of Different Dermatological Conditions Using Naïve Bayesian Classification”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 1, pp.864-868, January 2014.

[7] Dr. S. Vijayarani, Mr.S. Dhayanand,” Liver Disease Prediction using SVM and Naive Bayes Algorithms”, International Journal of Science, Engineering and Technology Research”, Volume 4, Issue 4, pp.816-820, April 2015.

[8] https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes.

[9] http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+ Wi sconsin+%28Original%29

[10] https://archive.ics.uci.edu/ml/datasets/Liver+Disorders

[11] https://archive.ics.uci.edu/ml/datasets/Statlog+(Heart)