# A COMPARATIVE STUDY OF CLASSIFICATION TECHNIQUES BY UTILIZING WEKA TOOL

[1]Manuja Sharma, [2]Dr. K.L. Bansal
[1]M.Tech Student, [2]Professor
Department of Computer Science,
Himachal Pradesh University, Shimla, India.

***Abstract:*** With the advent of technology, the amount of raw data which is available over the network is exploding day by day. It becomes extremely necessary to analyze this data and extract every possible information from it which helps in serving some or the other purpose. The way this huge amount of data is handled, analyzed and processed, this has become an interesting and fast growing area for research. Data mining offers various techniques and methods that may be used to predict the future trends and patterns in available data. There are number of techniques which can be used to predict the class in which the particular data falls. This research focuses on comparative analysis of various classification techniques, such as Naïve Bayes, Bayes Net, Random Forest, J48, Decision Table and lBK, which are used to predict the category of data. WEKA tool has been used to practically implement these classification techniques and the comparison has been done over the set of parameters like correctly and incorrectly classified instances, errors, kappa statistics, sensitivity, accuracy, specificity etc.

***Keywords*- Data Mining, Classification, Naïve Bayes, Bayes Net, Random Forest, J48, Decision Table, lBK, WEKA Tool.**

## I. INTRODUCTION

The amount of raw data is exploding day by day. This raw data is by itself of no use and does not provide much information if it remains unprocessed. In order to improve the utilization of data, various meaningful trends and patterns are extracted from this data with the help of data mining techniques. Data mining is the process of extracting useful information from the huge amount of data. It could also be defined as the process of discovering hidden trends and patterns from existing data and then using these trends and patterns to predict some future trends [1]. Numbers of data mining tools are available to analyze the data like WEKA, Knime, and Rapid Miner etc. These tools provide us with a collection of methods and techniques that are used to analyze the data in better ways. WEKA tool has been used in this research to analyze various classification techniques.

### 1.1 Classification

Classification is basically a data analysis task where this model is used to predict the class of data objects whose class labels are yet unknown. It is also defined as the process which helps in analyzing a set of data by generating some grouping rules which are further used to classify the future data. This is a two phase process:

1. In the learning step, a classifier or a classification model is built by describing some predetermined set of data classes or concepts. This is also known as the training phase where the algorithm builds the model by analyzing or learning from a training set made up of database tuples and the associated class labels [2].

2. In the classification step, the model extracted from the learning phase is then tested with the whole new test data set in order to measure and analyze the performance of trained model. If the performance measures are acceptable then the rule or the model is ready to be applied to new data tuples [2].

Following classification techniques have been focused in this paper:

#### 1.1.1 Naïve Bayes

Naïve Bayes classifier is a supervised learning technique which is based on the Bayes theorem. This technique is used in complex situations that deal with large data sets [3].

#### 1.1.2 Bayes Net

Bayes Net is the base class classifier in Bayesian Network techniques that involves learning using various search algorithms and quality measures [4].

#### 1.1.3 Random Forest

Random Forest is a multiple learning classifier which operates by constructing a multitude of decision trees at the time of training. This classifier helps to correct the problem of overfitting in decision trees at the time of training [5].

#### 1.1.4 J48

J48 is an extension of ID3 algorithm and an open source implementation of C4.5 algorithm. This algorithm tends to generate rules for predicting the class of target variable. Some of the additional features of J48 are accounting for missing values, pruning decision trees, rule derivation, continuous attribute value ranges, etc. [6].

### 1.1.5 Decision Table

Decision Table is a method which is used to numerically predict the data from decision tree. Decision table is a rule based classifier which is an ordered set of IF-THEN rules that are much more compact and are much easier to comprehend than that of decision trees [7].

### 1.1.6 lBK

lBK is k-nearest neighbor classifier. It can select the appropriate value of k based on cross-validation and is also capable of doing distance weighing [8].

## II. RELATED WORK

Some of the important works where the classification algorithms have been comparatively analyzed are:

**Nazum Nahar, Ferdous Ara (2018)** presented a study that explored the early prediction of liver disease by using various decision tree classification techniques. WEKA tool was used to calculate and compare the performance of seven classifiers, namely J48, LMT, Random Forest, REP Tree, Decision Stump, Random Tree and Hoeffding Tree. The results showed that the Decision Stump outperformed all the other classifiers [9].

**Anand Kishore Pandey, Dharmveer Singh Rajpoot (2016)** comparatively analyzed some classification algorithms using WEKA tool. The analysis was done on the dataset of alcohol consumption by school students. The comparison was done among six algorithms, namely Decision Stump, Random Forest, J48, Naïve Bayes, Naïve Bayes Simple and Bayes Net. As a result it was showed that among all these mentioned algorithms, Decision Stump approach performed the better classification [10].

**G V Gayathri, B Siva Jyothi (2018)** conducted a study to experimentally analyze some data mining techniques in an attempt to find the most suitable classifier to categorize text messages as spam and non-spam. The performance of five classifiers was checked, namely Naïve Bayes, SVM, Logistics Model, Decision Tree and Random Forest classifier. The experimental results showed that the performance of Random Forest classifier outperformed all the above mentioned classifiers [11].

**Vikas Chaurasia, Saurabh Pal (2017)** conducted a study with the aim of investigating the performance of different classification algorithms to detect the breast cancer disease in women at an early stage. Three classifiers, namely SMO, lBK and BF Tree were comparatively analyzed and implemented over WEKA environment. It was concluded that the SMO has higher prediction accuracy and outperformed the other algorithms [12].

**Poonam Rani, Navpreet Rupal (2018)** used WEKA tool to analyze the traffic data with the help of classification algorithms. For predicting the traffic and analyzing the performance, four classifiers were compared, namely J48, Random Forest, Decision Tree and Naïve Bayes. It was concluded that Random Forest outperformed all the other classifiers and performs best with traffic data [13].

**Mrs. T. Seeni Selvi (2018)** used WEKA tool to classify the agricultural land soils of different states in India. Two classification algorithms were used in this study, Random Forest and Decision Stump. These algorithms were evaluated over the India Crop Production State-wise dataset on the basis of one single parameter, i.e. accuracy. The results showed that Random Forest outperformed Decision Stump with 99% accuracy rate [14].

## III. WEKA TOOL

WEKA stands for Waikato Environment for Knowledge Analysis. WEKA tool is the most important tool in data mining. WEKA uses a collection of machine learning algorithms, developed by The University of Waikato, Hamilton, New Zealand. These algorithms can be applied directly to the data or data called from the Java code. WEKA is free software licensed under the GNU General Public License. WEKA uses ARFF (Attribute Relation File Format) file for the analysis of data, by default. Other file formats like CSV (Comma Separated Values), C4.5 data files etc. and databases using ODBC, from where data can be imported. WEKA is a collection of algorithms for: Classification, Regression, Clustering, Association, Data pre-processing and Visualization [15].

## IV. RESULTS AND DISCUSSION

For the practical implementation of this work, WEKA tool has been used to implement the following classification algorithms: Naïve Bayes, Bayes Net, Random Forest, J48, Decision Table and lBK. Overfitting is avoided by using the 10-fold cross validation method. The Adult data set has been taken into consideration from the UCI machine learning repository. This is a multivariate data set which comprises of 14 attributes (categorical and integral), 48842 number of instances with some missing values. The prediction task is to determine whether a person's income exceed 50K a year or not based on census data [16].

These algorithms are analyzed and evaluated on the basis of parameters:

Correctly and incorrectly classified instances, kappa statistics, errors, precision, recall, F-measure, TP-rate, FP-rate, accuracy, specificity and sensitivity.

### 4.1 Correctly and Incorrectly Classified Instances

From Table 4.1, it can be concluded that the percentage of correctly classified instances by these classification algorithms is more than the percentage of incorrectly classified instances. The results show that J48 algorithm correctly classifies larger number of instances with respect to other algorithms.

Figure 4.1 shows the bar graph for the correctly and incorrectly identified instances by these algorithms.

Table 4.1: Correctly and Incorrectly Classified Instances among Various Algorithms.

| Algorithms | Correctly Classified Instances | Incorrectly Classified Instances |
|---|---|---|
| Naïve Bayes | 83.4889% | 16.5111% |
| Bayes Net | 83.8913% | 16.1087% |
| Random Forest | 84.7082% | 15.2918% |
| J48 | 86.1118% | 13.8882% |
| Decision Table | 85.6665% | 14.3335% |
| lBK | 79.2721% | 20.7279% |

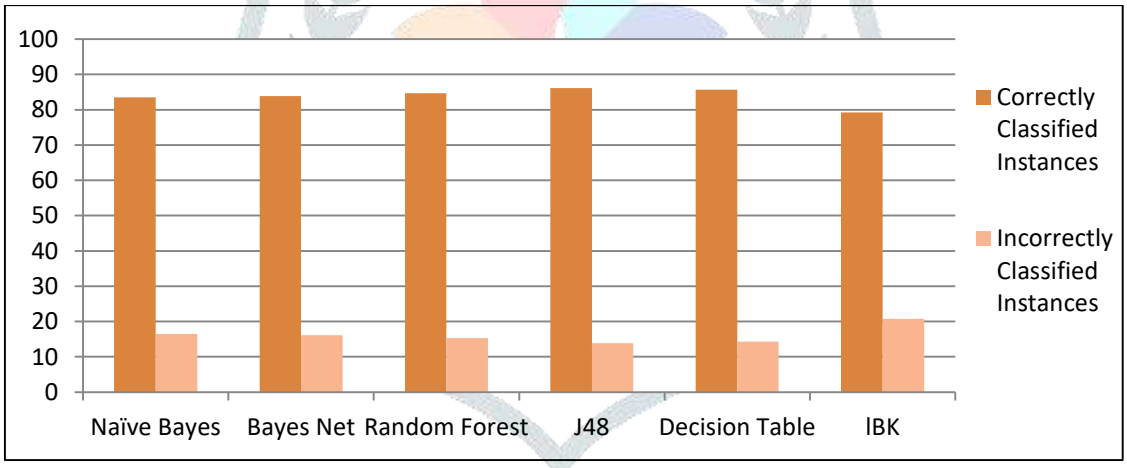


Fig. 4.1: Graph To Represent Correctly and Incorrectly Classified Instances.

### 4.2 Kappa Statistics

Kappa refers to a chance-corrected measure which is calculated between classification and true classes. Such a measure is computed by taking the expected attribute from the observed values of attributes. The value is then divided by the maximum value of the attribute. Value greater than zero indicates a better performance as compared to chance. In the case of our data set, J48 performs better with respect to other algorithms. Table 4.2 and Figure 4.2 show the results.

Table 4.2: Kappa Statistics for Various Algorithms.

| Algorithms | Kappa Statistics |
|---|---|
| Naïve Bayes | 0.5024 |

| | |
|---|---|
| Bayes Net | 0.5961 |
| Random Forest | 0.5634 |
| J48 | 0.5988 |
| Decision Table | 0.5662 |
| lBK | 0.4281 |

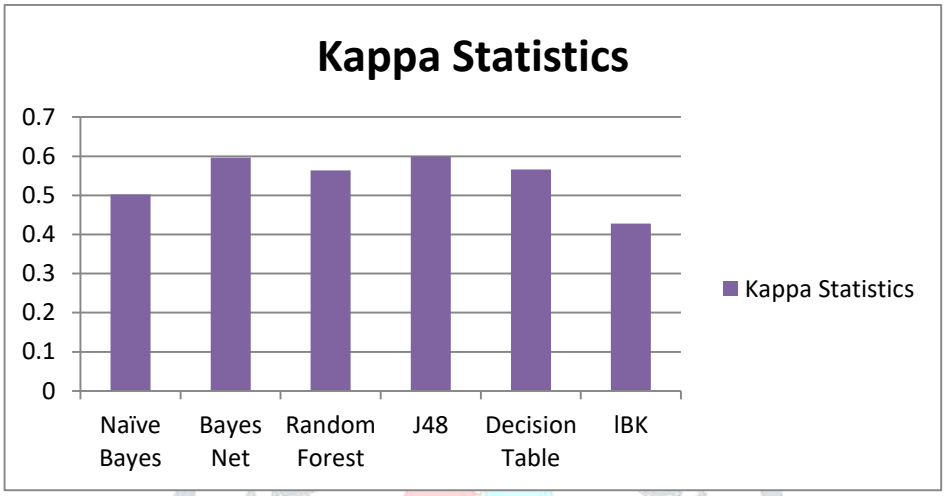


Fig. 4.2: Graph to Represent the Kappa Statistics Values.

**4.3 Errors**

Mean Absolute Error (MAE) computes the average magnitude of errors. Root Mean Squared Error (RMSE) also computes the average magnitude of errors, but the difference lies in the way that the difference between the predicted and absolute observation is squared and is then averaged over the set of observations. The results in Table 4.3 and Figure 4.3 show that Naïve Bayes performs better on this parameter with respect to other algorithms.

Table 4.3: MAE and RSME Values for Various Algorithms.

| **Algorithms** | **Mean Absolute Error (MAE)** | **Root Mean Squared Error (RMSE)** |
|---|---|---|
| Naïve Bayes | 0.1731 | 0.3716 |
| Bayes Net | 0.1759 | 0.343 |
| Random Forest | 0.1975 | 0.3271 |
| J48 | 0.1925 | 0.3216 |
| Decision Table | 0.2072 | 0.3186 |
| lBK | 0.2073 | 0.4553 |

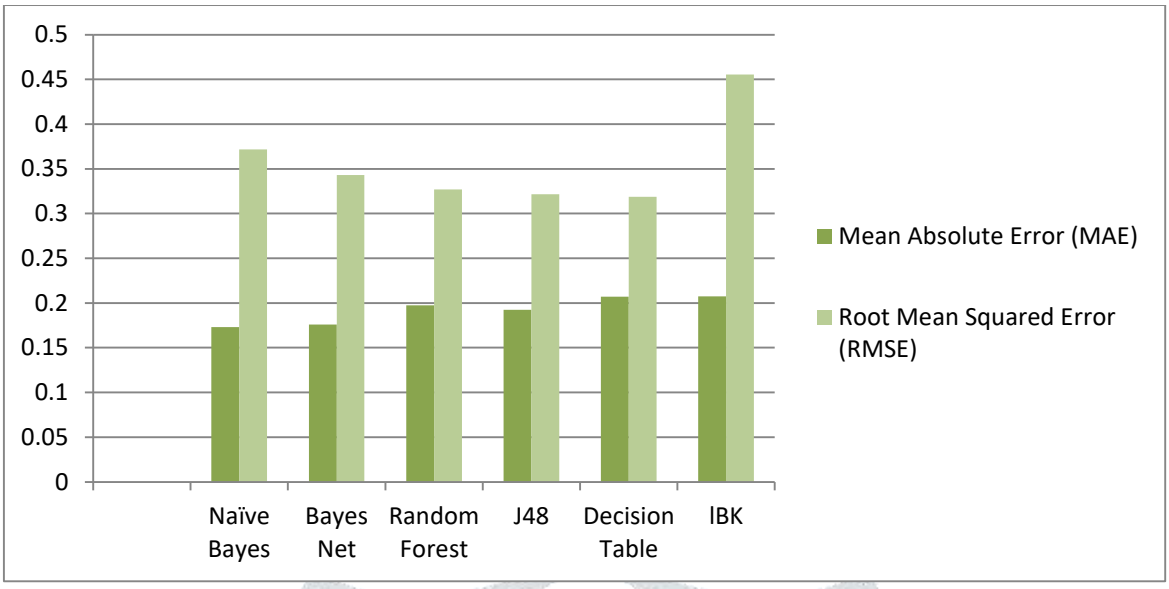


Fig. 4.3: Graph to Represent MAE and RSME values.

Relative Absolute Error (RAE) and Root Relative Squared Error are those errors with which the performance of every experiment is computed. Absolute error gives us the amount of physical error, while relative error provides us with the information about how much efficient a particular measurement is with respect to size of the attribute being measured. Table 4.4 and Figure 4.4 shows that lBk gives a better performance over this parameter.

Table 4.4: RAE and RRSE Values for Various Algorithms.

| **Algorithms** | **Relative Absolute Error (RAE)** | **Root Relative Squared Error (RRSE)** |
|---|---|---|
| Naïve Bayes | 47.3347% | 86.9042% |
| Bayes Net | 48.1086% | 80.2261% |
| Random Forest | 54.0053% | 76.5111% |
| J48 | 52.6509% | 75.2037% |
| Decision Table | 56.6655% | 74.518% |
| lBK | 56.6963% | 106.4726% |

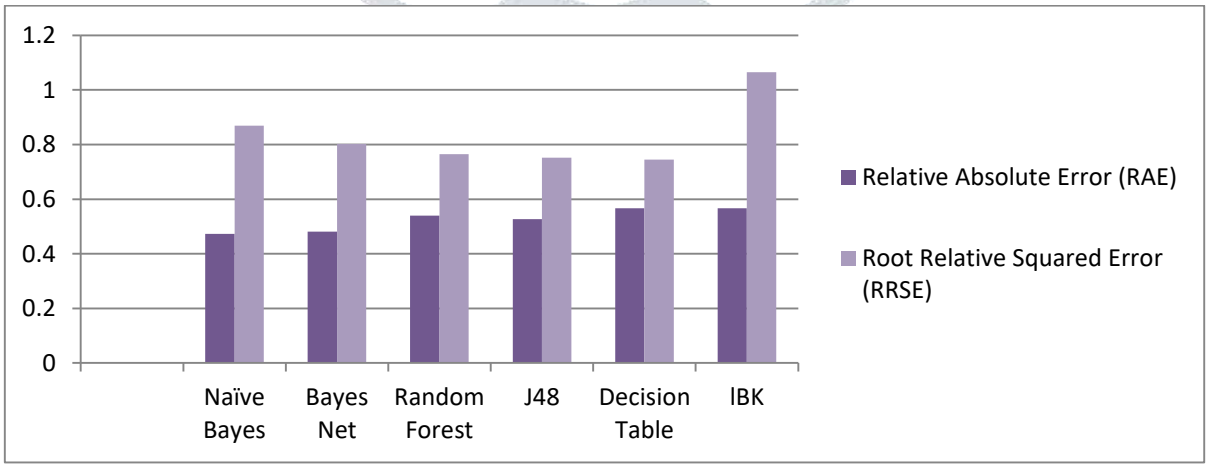


Fig. 4.4: Graph to Represent RAE and RRSE values.

**4.4 Accuracy Measures**

The accuracy of the classification algorithms is measures with the help of parameters such as TP-rate, FP-rate, precision, recall, F-measure. These parameters are defined as:

1. TP-rate: It is known as the rate of True Positives. TP-rate defines the instances that have been correctly classified with respect to the given class.

2. FP-rate: It is known as the rate of False Positives. FP-rate defines the instances that have been incorrectly classified with respect to the given class.

3. Precision: This parameter lists the proportion of those instances which are true to a particular class divided by overall instances classified with respect to that class.

4. Recall: This measure defines the proportion of those instances that have been classified by a class divided by the total instances present in the class.

5. F-measure: It is calculated as:

$$Measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

Table 4.5 and Figure 4.5 show the values of the parameters mentioned above.

Table 4.5: Values of TP-rate, FP-rate, Precision, Recall and F-measure for various algorithms.

| Algorithms | TP-rate | FP-rate | Precision | Recall | F-measure |
|---|---|---|---|---|---|
| Naïve Bayes | 0.835 | 0.379 | 0.825 | 0.835 | 0.825 |
| Bayes Net | 0.839 | 0.189 | 0.858 | 0.839 | 0.845 |
| Random Forest | 0.847 | 0.307 | 0.842 | 0.847 | 0.844 |
| J48 | 0.861 | 0.294 | 0.856 | 0.861 | 0.857 |
| Decision Table | 0.857 | 0.345 | 0.851 | 0.857 | 0.848 |
| lBK | 0.793 | 0.368 | 0.791 | 0.793 | 0.792 |

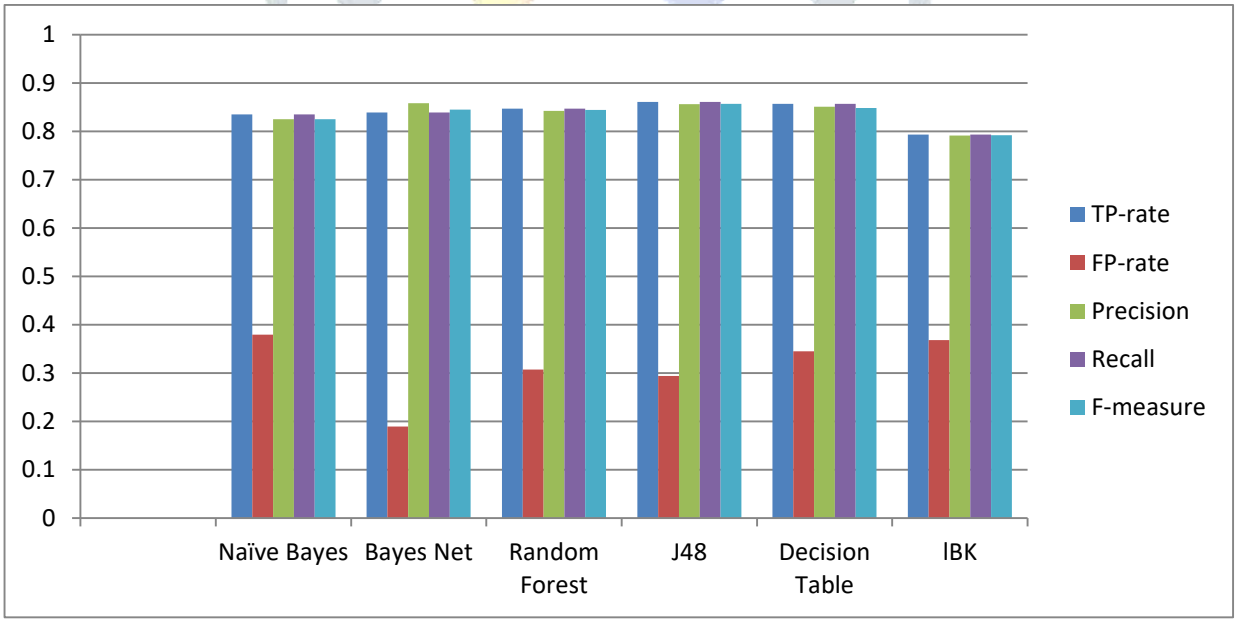


Fig. 4.5: Graph to Represent the Values of Accuracy Parameters.

*Specificity*, *sensitivity* and *accuracy* of the classification algorithms is obtained using confusion matrix.

Table 4.6: Confusion Matrix.

| | Correctly Classified | Incorrectly Classified |
|---|---|---|
| **Selected** | TP | FP |
| **Not Selected** | FN | TN |

Here, TP: True Positive
FP: False Positive
TN: True Negative
FN: False Negative

Specificity, sensitivity and accuracy of the algorithms are calculated by the formula:

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

The overall performance of the classification algorithms is calculated using these three measures, i.e. specificity, sensitivity and accuracy. Table 4.7 and Figure 4.6 show the values of these measures and it is thus concluded that the accuracy of J48 algorithm is highest among all the other classification algorithms analyzed here.

Table 4.7: Comparison of Specificity, Sensitivity and Accuracy.

| Algorithms | Specificity | Sensitivity | Accuracy |
|---|---|---|---|
| Naïve Bayes | 0.7156 | 0.1654 | 0.8348 |
| Bayes Net | 0.6307 | 0.2532 | 0.8389 |
| Random Forest | 0.7081 | 0.1969 | 0.8471 |
| J48 | 0.7508 | 0.2011 | 0.8611 |
| Decision Table | 0.7821 | 0.1780 | 0.8567 |
| lBK | 0.5924 | 0.1592 | 0.7927 |

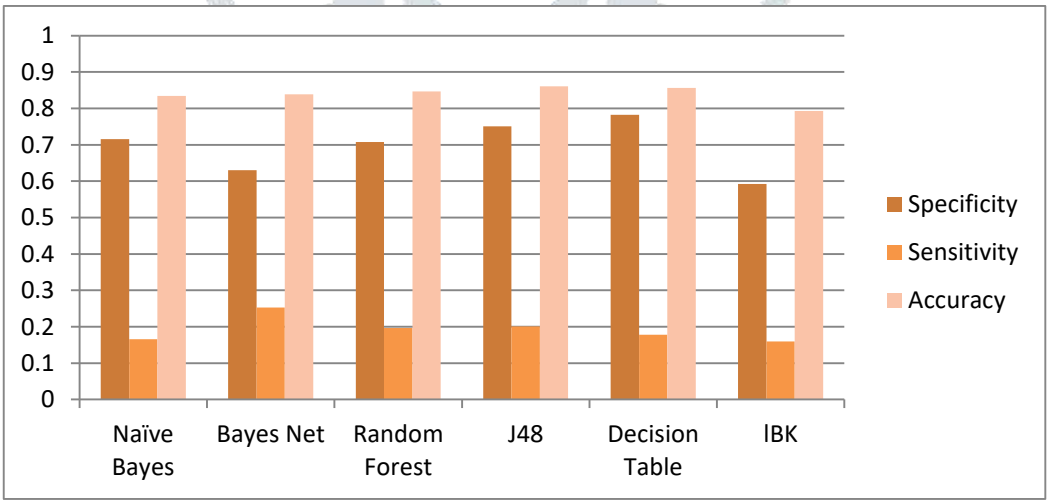


Fig. 4.6: Graph to Represent the Values of Specificity, Sensitivity and Accuracy.

From these comparisons, it is concluded that different algorithms perform differently over the particular set of parameters and there is no single algorithm which would perform on the same accuracy rate on every type of data. Therefore, there is no free-lunch policy in case of classification algorithms. The performance of algorithms depends highly on the type of the data used.

## V. CONCLUSION AND FUTURE SCOPE

In this work, the performance of six classification algorithms is analyzed, i.e. Naïve Bayes, Bayes Net, Random Forest, J48, Decision Table and lBK. There are so many benchmarks for comparing the performance and accuracy of these classification algorithms. All these algorithms are compared on the basis of parameters like correctly and incorrectly classified instances, kappa statistics, mean absolute error, root mean squared error, relative absolute error, root relative squared error, TP-rate, FP-rate, precision, recall, F-measure, specificity, sensitivity and accuracy. It has been observed form the experiments that J48 performed better on the parameters of correctly and incorrectly classified instances and kappa statistics. Naïve Bayes resulted better in terms of MAE and RSME values while lBK performed better in terms of RAE and RRSE Values. From the results, it has been concluded that overall performance of J48 algorithm is better and J48 has outperformed all the other algorithms in terms of accuracy i.e. J48 has 86% accuracy. Therefore, there is no free-lunch policy in case of classification algorithms. The performance of algorithms depends highly on the type of the data used.

For the future scope, same algorithms can be implemented on different data over some different application domain or tool instead of WEKA and their performance can be analyzed and improved with some multiple learning techniques with the help of some different tools.

## REFERENCES


[1] Jiawei Han, Micheline Kamber and Jian Pei, "Data Mining: Concepts and Techniques, Morgan Kaufmann", Third Edition, 2011.

[2] Sagar S. Nikam, "A Comparative Study of Classification Techniques in Data Mining Algorithms", Oriental Journal of Computer Science and Technology, Vol 8(1), pp 13-19,2015.

[3] Hetal Bhavsar, Amit Ganatra, "A Comparative Study of Training Algorithms for Supervised Machine Learning", International Journal of Soft Computing and Engineering (IJSCE), Volume-2, Issue-4, September 2012.

[4] Available at: https://www.cs.waikato.ac.nz/~remco/weka.bn.pdf

[5] Available at: https://en.wikipedia.org/wiki/Random_forest

[6] V. Krishnaiah, Dr. G. Narsimha and Dr. N. Subhash Chandra, "Survey of Classification Techniques in Data Mining", International Journal of Computer Sciences and Engineering, Vol 2(9), pp 65-74, 2014.

[7] G. Kesavaraj and S. Sukumaran, "A Study On Classification Techniques in Data Mining", ICCCNT Fourth International Conference on IEEE, pp 1-7,2013.

[8] T. Augustine, P. Vasudeva Reddy, P.V.G.D. Prasad Reddy, "A Framework for Performance Evaluation Of Classifiers: Case Study on NIDS", International Journal of Pure and Applied Mathematics, 2018.

[9] Nazmun Nahar and Ferdous Ara, "Liver Disease Prediction By Using Different Decision Tree Techniques", International Journal of Data Mining and Knowledge Management Process, Vol.8, No.2, March 2018.

[10] Anand Kishor Pandey and Dharmveer Singh Rajpoot, "A comparative study of classification techniques by utilizing WEKA", IEEE, 2016.

[11] G V Gayathri and B Siva Jyothi, "A Comparative Study of Classification Algorithms on Spam Detection", International Journal for Research in Applied Science & Engineering Technology, Volume 6 Issue IV, April 2018.

[12] Vikas Chaurasia and Saurabh Pal, "A Novel Approach for Breast Cancer Detection using Data Mining Techniques", 2017.

[13] Poonam Rani, Navpreet Rupal, "Traffic Data Analysis Using Decision Tree And Naïve Bayes Classifier", International Research Journal of Engineering and Technology (IRJET), Volume: 05 Issue: 05, May-2018.

[14] Mrs. T. Seeni Selvi, "Efficient Classification of Agriculture Land Soils In State-wise From India Using Data Mining With Weka", International Journal of Scientific Research in Computer Science, Engineering and Information Technology, Volume 3,Issue 3, ISSN : 2456-3307.

[15] WEKA, the University of Waikato, Available at: http://www.cs.waikato.ac.nz/ml/weka/

[16] Data set from UCI Machine Learning Repository, Available at: https://archive.ics.uci.edu/ml/datasets/Adult