*International Journal of Research in Advent Technology, Vol.7, No.5S, May 2019*
*E-ISSN: 2321-9637*
*Available online at www.ijrat.org*

# A Comparative Study on Machine Learning with Ensemble Learning for Predicting Students' Academic Performance In Educational Data Mining

V. Vanthana,
*Computer Applications Department,*
*The Standard Fireworks Rajaratnam College For Women, Sivakasi*
*vanthana.310@gmail.com*

**Abstract -** Now a days many higher education institutions prefer data mining and machine learning techniques to analyse the academic improvement of their students. To sustenance that the concept of Educational Data Mining arises. Also, to improve the performance of the existing machine learning algorithms, ensemble learning methods are used. This paper proposes a comparative study on the two machine learning algorithms – Decision Tree and Naïve Bayes with three ensemble learning methods – Bagging, Boosting and Voting. The algorithms are applied on the data collected from three colleges of Assam, India. The data consists of socio-economic, demographic as well as academic information of three hundred and fifty students with twenty-two attributes. The data mining tool used was WEKA. The data mining task here is to predict the students' performance in the final semester. The obtained result reveals that Adaboost greatly improves the accuracy of Naïve Bayes than Bagging and Bagging greatly improves the accuracy of Decision Tree than Adaboost. Also Voting provides the accuracy nearly within the range of accuracy of the decision tree in bagging and boosting.

**Keywords -** Educational Data Mining, Students Academic Performance, WEKA, Machine learning, ensemble learning, Bagging, Boosting, Voting.

## 1. INTRODUCTION

The implementation of data mining in the educational sector, recently defined as "educational data mining" (EDM) [1], is a new stream in the data mining research field. The educational data mining research community is constantly growing, starting by organizing workshops since 2004, then conducting an annual International Conference on EDM beginning since 2008, and now already having a Journal on EDM (the first issue being published in October 2009). There are already a large number of research papers discussing various problems within the higher education sector and providing examples for successful solutions reached by using data mining. Also there are many research papers that analyze the data collected by the institutions using many data mining algorithms.

Students' academic performance was assessed to predict the final exam result by Sadiq Hussain, Neama Abdulaziz Dahan, Fadl Mutaher Ba-Alwi, Najoua Ribata in 2018 [2]. Modern educational institutes need data mining for their strategy and future plans. Student's performance depends on
various factors like personal, social, economic and other environmental ones [3,4]. The top-level educational institutes' authorities may utilize the outcome of the experimental results to understand the trends and behaviors in students' performance which may lead to design new pedagogical strategies [5].

There are a number of machine learning and ensemble learning algorithms: Decision Tree, Neural Network, Naïve Bayes, KNearest neighbor, Random Forest, AdaBoost, Support Vector Machines etc. [6]. In this research, two machine learning algorithms are going to be used for mining the academic students' performance: Decision Tree, Naïve Bayesian and the three ensemble learning methods: Bagging, Boosting and Voting are used to improve the performance of them. In this study, students' end semester percentage is selected as the dependent parameter. The percentages are categorized as 'Best', 'Very Good', 'Good', 'Pass', 'Fail'. The data mining tool used for the study was WEKA, which is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization.

## 2. RELATED WORK

Predicting students' performance is an important task in the web based educational models. To build a predictive model, there are many Data Mining techniques used like classification, regression and clustering. The most popular technique to predict students' performance is classification. There are several methods under classification such as Decision tree, Naïve Bayes, etc.

Superby et al. [7] predict students at risk of drop-out, determining factors influencing the achievement of the first-year university students, classifying students into three

classes – low-risk, medium-risk and high-risk, using Decision trees, Random forest method, Neural networks and Linear discriminant analysis. Vandamme et al. [8] also deals with early identification of three categories of students: low, medium and high-risk students using Decision trees, Neural networks and Linear discriminant analysis.

Cortez and Silva in [9] attempt to predict student failure by applying and comparing four data mining algorithms, Decision Tree, Random Forest, Neural Network and Support Vector Machine. The implementation of predictive modelling for maximizing student recruitment and retention is presented in the study of Noel-Levitz [10].

Ajay Kumar Pal and Saurabh Pal collected the data for the study and analysis of the student's educational performance basically for training and placement. The authors used different classification algorithm and used WEKA data mining tool [11]. They concluded that naive Bayes classification model is the better algorithm based on the placement data with found accuracy of 86.15% and overall time taken to build the model is at 0 sec. As compared with others Naïve Bayes classifier had lowest average error i.e. 0.28

## 3. EVALUATION MEASURES FOR THE CLASSIFIERS

A binary classifier predicts all data instances of a test dataset as either positive or negative. This classification (or prediction) produces four outcomes – true positive, true negative, false positive and false negative.

- True positive (TP): correct positive prediction
- False positive (FP): incorrect positive prediction
- True negative (TN): correct negative prediction
- False negative (FN): incorrect negative prediction

### *A. Confusion Matrix*

It is two by two table constructed by counting the number of the four outcomes of the classifier.

| Predicted Vs Observed | **Positive** | **Negative** |
|---|---|---|
| **Positive** | TP | FN |
| **Negative** | FP | TN |

### *B. Measures Derived From The Confusion Matrix*

Various measures can be derived from a confusion matrix.

*1) Accuracy*

Accuracy (ACC) is the number of all correct predictions divided by the total number of the dataset. The best accuracy is 1.0, whereas the worst is 0.0.

$$\text{Accuracy} = (TP + TN) / (P + N)$$

*2) Sensitivity*

It is also called as Recall or True positive rate. It is calculated as the number of correct positive predictions divided by the total number of positives. The best sensitivity is 1.0, whereas the worst is 0.0.

$$\text{Sensitivity} = (TP / P)$$

*3) Precision*

It is the number of correct positive predictions divided by the total number of positive predictions. It is also called positive predictive value (PPV). The best precision is 1.0, whereas the worst is 0.0.

$$\text{Precision} = TP / (TP + FP)$$

*4) False Positive rate*

It is the number of incorrect positive predictions divided by the total number of negatives. The best false positive rate is 0.0 whereas the worst is 1.0. It can also be calculated as 1 – specificity.

$$FPR = (FP / N)$$

*5) F-score*

F-score is a harmonic mean of precision and recall.

$$F= 2PR / (P+R)$$

## 4. DATA COLLECTION

The data set contains 350 instances with 22 attributes which is collected from the UC machine learning datasets. It encloses the details of the students collected from the three colleges of Assam in India. The features are as follows

| **Feature** | **Description** | **Domain Values** |
|---|---|---|
| ge | Gender | M, F |
| cst | Caste | G, ST, SC, OBC, MDBC |
| tnp | Class X % | Best, Vg, Good, Pass, Fail |
| twp | Class XII % | Best, Vg, Good, Pass, Fail |
| iap | Internal assessment percentage | Best, Vg, Good, Pass, Fail |
| esp | External assessment percentage | Best, Vg, Good, Pass, Fail |
| arr | Arrear or not | Y, N |
| ms | Marital status | Married, Unmarried |
| ls | Living in Town or Village | T, V |
| as | Admission category | Free, Paid |
| fmi | Family Monthly Income | Vh, High, Medium, Low |
| Fs | Family Size | Large, Average, Small |
| Fq | Father Qualification | Nil, 10, 12, Degree, PG |
| Mq | Mother Qualification | Nil, 10, 12, Degree, PG |
| Fo | Father Occupation | Service, Business, Retired, Farmer, Others |
| Mo | Mother Occupation | Service, Business, Retired, House wife, Others |
| Nf | Number of Friends | Large, Average, Small |
| Sh | Study hours | Good, Average, Poor |

| | | |
|---|---|---|
| Ss | School studied | Govt, Private |
| Me | Medium | Eng, Asm, Hin, Ben |
| Tt | Home to college Travel Time | Large, Average, Small |
| Atd | Class Attendance Percentage | Good, Average, Poor |

## 5. DATA PREPROCESSING

### *A) Data Cleaning*

Data Cleaning is one of the preprocessing tasks. It is used to eliminate unwanted and missing values. Here in this data set there are 30 instances with missing values. So they are removed and only 320 instances are used for classification

### *B) Feature Selection*

The Feature Selection is the most important Preprocessing task. The objective is to select the appropriate subset of features which can better describes the data, removes irrelevant data and reduces the dimensionality of the feature space. [12]. There are two categories of Feature Selection methods – wrapper based and filter based. Filter based method means searching for the minimum set of relevant features while ignoring the rest. It uses the ranking techniques to rank the features and only the highly ranked features are selected for the algorithm. There are many ranking techniques like information gain, gain ratio, gini index, annova, etc.

In this paper, filter method based on the information gain is used to select the relevant features. Here each feature is assigned ranking based on their influence on classification. The highly ranked features are selected while others are excluded. Here only the 12 out of 22 got the high rank and selected as the relevant features. They are as follows,

*Atd, ms, ls, fmi, arr, iap, twp, tnp, cst, as, fs, tt*

## 6. PROPOSED STUDY ON MACHINE LEARNING WITH ENSEMBLE LEARNING

### *A. Methodology*

In this research, the two machine learning algorithms – Decision Tree(J48), Naïve Bayes are applied on the data collected from three colleges of Assam, India and the classifier is build. These classifiers are evaluated using confusion matrix measures. Then the three ensemble methods – Bagging, Boosting and Voting are applied to these two machine learning algorithms and the classifier is evaluated. Then the evaluation results are studied comparatively. Here for the training 66% of the tuples are taken and the remaining are used for Testing.

### *B. Application Of Machine Learning Algorithms To The Dataset*

*1) Decision tree(J48)*

Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. Decision tree J48 is the implementation of algorithm ID3 developed by the WEKA project team. Here J48 is applied to the dataset and their evaluation measures are as follows

| Class Value | TP rate | FP rate | Precision | Recall | F-score |
|---|---|---|---|---|---|
| **Good** | 0.600 | 0.229 | 0.522 | 0.600 | 0.558 |
| **Vg** | 0.733 | 0.316 | 0.647 | 0.733 | 0.688 |
| **Best** | 0.333 | 0.032 | 0.500 | 0.333 | 0.400 |
| **Pass** | 0.417 | 0.036 | 0.714 | 0.417 | 0.526 |
| **Average** | 0.603 | 0.216 | 0.609 | 0.603 | 0.596 |
| **Accuracy** : 60.2941% | | | | | |

*Table 1 Accuracy of J48*

Here for the J48, all the evaluation measures provided approximately the best result. The accuracy calculated id 60.2941%.

*2) Naïve Bayes*

Next the data set is applied to the Naïve Bayesian algorithm. In machine learning, naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

Naive Bayes has been studied extensively since the 1960s. It was introduced (though not under that name) into the text retrieval community in the early 1960s,[1] and remains a popular (baseline) method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features. Their evaluation measures using WEKA is as follows

| Class Value | TP rate | FP rate | Precision | Recall | F-score |
|---|---|---|---|---|---|
| **Good** | 0.900 | 0.271 | 0.581 | 0.900 | 0.706 |
| **Vg** | 0.700 | 0.053 | 0.913 | 0.700 | 0.792 |
| **Best** | 0.500 | 0.016 | 0.750 | 0.500 | 0.600 |
| **Pass** | 0.667 | 0.036 | 0.800 | 0.667 | 0.727 |
| **Average** | 0.735 | 0.111 | 0.781 | 0.735 | 0.739 |
| **Accuracy** : 73.5294% | | | | | |

*Table 2 Accuracy of Naïve Bayes*

In Naïve Bayes, the evaluation results are much better that the J48 and the accuracy is 73.5294%.

## 7. USING ENSEMBLE LEARNING METHODS

In this paper, three ensemble learning methods – Bagging, Boosting and Voting are applied to the above two algorithms and the classifier is evaluated.

***A. Bagging***

Bootstrap aggregating, also called bagging, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting. Although it is usually applied to decision tree methods, it can be used with any type of method. Bagging is a special case of the model averaging approach.

All the Bootstrap samples will be trained using different classifiers. Individual classifiers are then combined through major vote process, the class chosen was by the most number of classifiers in the ensemble design [13]

***B. Boosting***

Boosting is a general ensemble method that creates a strong classifier from a number of weak classifiers. This is done by building a model from the training data, then creating a second model that attempts to correct the errors from the first model. Models are added until the training set is predicted perfectly or a maximum number of models are added.

AdaBoost was the first really successful boosting algorithm developed for binary classification. It is the best starting point for understanding boosting. Modern boosting methods build on AdaBoost, most notably stochastic gradient boosting machines. In this research, Adaboost algorithm has been used.

***C. Voting*** Voting is one of the simplest ways of combining the predictions from multiple machine learning algorithms. It works by first creating two or more standalone models from the training dataset. A Voting Classifier can then be used to wrap your models and average the predictions of the sub-models when asked to make predictions for new data.

The predictions of the sub-models can be weighted, but specifying the weights for classifiers manually or even heuristically is difficult. It is possible to create a voting ensemble model for classification using the VotingClassifier class.In this research two classifiers – Tree, Naïve Bayes are calculated and voted based on the average of probability

***D. Experimental Results***

*1) Accuracy of J4.8 using Ensemble Methods*

| Class Value | TP rate | FP rate | Precision | Recall | F-score |
|---|---|---|---|---|---|
| **Good** | 0.700 | 0.146 | 0.667 | 0.700 | 0.683 |
| **Vg** | 0.733 | 0.211 | 0.733 | 0.733 | 0.733 |
| **Best** | 0.333 | 0.016 | 0.667 | 0.333 | 0.444 |
| **Pass** | 0.750 | 0.089 | 0.643 | 0.750 | 0.692 |
| **Average** | 0.991 | 0.153 | 0.692 | 0.691 | 0.686 |
| **Accuracy** : 69.1176% | | | | | |

*Table 1 J48 using Bagging*

All the evaluations measures are improved while applying bagging on J48. The accuracy improves from 60.2941% to 69.1176%.

| Class Value | TP rate | FP rate | Precision | Recall | F-score |
|---|---|---|---|---|---|
| **Good** | 0.800 | 0.188 | 0.640 | 0.800 | 0.711 |
| **Vg** | 0.700 | 0.184 | 0.750 | 0.700 | 0.724 |
| **Best** | 0.500 | 0.032 | 0.600 | 0.500 | 0.545 |
| **Pass** | 0.667 | 0.036 | 0.800 | 0.667 | 0.727 |
| **Average** | 0.706 | 0.146 | 0.713 | 0.706 | 0.705 |
| **Accuracy** : 70.5882% | | | | | |

*Table 2 J48 using Adaboost*

*Here while using Boosting (AdaBoost) on J48 the accuracy improves from 60.2941% to 70.5882%.*

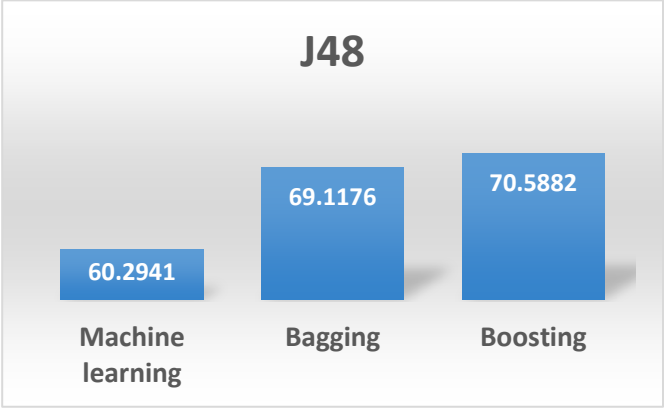


*Fig 1. Accuracy variations of J48*

The proposed result reveals that the J48 will offer better efficiency on Boosting than Bagging.

*2) Accuracy of Naïve Bayes using Ensemble Methods*

| Class Value | TP rate | FP rate | Precision | Recall | F-score |
|---|---|---|---|---|---|
| **Good** | 0.750 | 0.167 | 0.652 | 0.750 | 0.698 |
| **Vg** | 0.767 | 0.132 | 0.821 | 0.767 | 0.793 |
| **Best** | 0.500 | 0.032 | 0.600 | 0.500 | 0.545 |
| **Pass** | 0.833 | 0.036 | 0.833 | 0.833 | 0.833 |
| **Average** | 0.750 | 0.116 | 0.754 | 0.750 | 0.634 |
| **Accuracy** : 75% | | | | | |

*Table 3 Accuracy using Bagging*

While applying Bagging on Naïve Bayes, the accuracy will improve to 75% from 73.5294.

| Class Value | TP rate | FP rate | Precision | Recall | F-score |
|---|---|---|---|---|---|
| Good | 0.900 | o.229 | 0.621 | 0.900 | 0.735 |
| Vg | 0.767 | 0.105 | 0.852 | 0.767 | 0.807 |
| Best | 0.333 | 0.032 | 0.500 | 0.333 | 0.400 |
| Pass | 0.500 | 0.036 | 0.750 | 0.500 | 0.600 |
| Average | 0.721 | 0.123 | 0.735 | 0.721 | 0.713 |
| **Accuracy** : 72.0588% | | | | | |

*Table 4 Accuracy using Boosting*

*The Boosting will drop the performance of Naïve Bayes from 73.5294% to 72.0588%.*

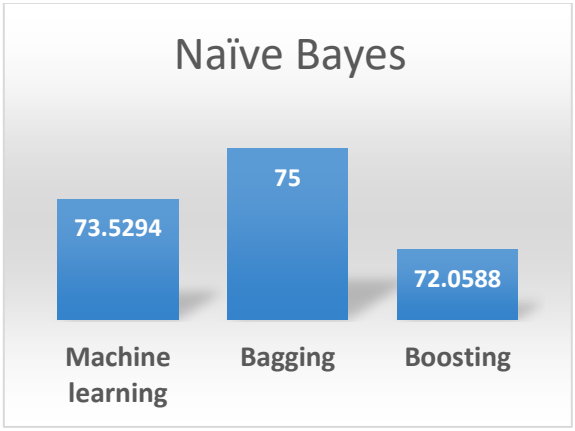


*Fig 2. Accuracy Variation of Naïve Bayes*

The proposed result reveals that the performance of Naïve Bayes will be improved while using Bagging. For Boosting the result will be approximately the same as that of the machine learning.

*3) J48 and Naïve Bayes for Voting*

| Class Value | TP rate | FP rate | Precision | Recall | F-score |
|---|---|---|---|---|---|
| Good | 0.600 | 0.208 | 0.545 | 0.600 | 0.571 |
| Vg | 0.733 | 0.289 | 0.667 | 0.733 | 0.698 |
| Best | 0.333 | 0.032 | 0.500 | 0.333 | 0.400 |
| Pass | 0.583 | 0.036 | 0.778 | 0.583 | 0.667 |
| Average | 0.632 | 0.198 | 0.636 | 0.632 | 0.629 |
| **Accuracy** : 63.2353% | | | | | |

*Table 5 Voting*

Here the two classifiers – Tree and Naive Bayes are combined and the average of probabilities are used for voting. The accuracy measure here is 63.2353% which lies between the accuracy of J48 (60.2941%) and the accuracy of Naïve Bayes (73.5294%).

***E. Accuracy of Machine learning with Ensemble learning***

| Algorithm | Without using Ensemble learning | Using Ensemble learning | | |
|---|---|---|---|---|
| | | Bagging | Boosting | Voting |
| J48 | 60.2941 | 69.1176 | 70.5882 | 63.2353 |
| Naïve Bayes | 73.5294 | 75 | 72.0588 | |

*Table 6 Comparative study of Accuracies in percentage*

The proposed comparative study reveals that all the three ensemble methods - bagging, boosting and Voting improves the accuracy of J48. But for Naïve Bayes only Bagging method improves the accuracy.

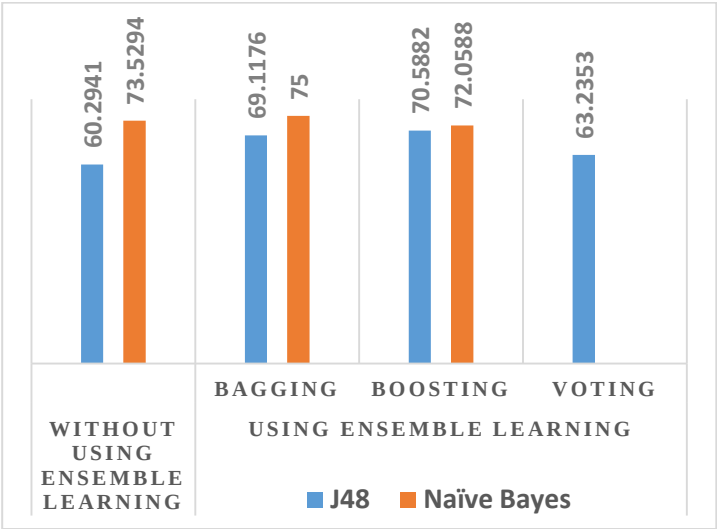


*Fig 3 Comparative study of Accuracies*

## 8. CONCLUSION AND FUTURE WORK

The academic performance of the students was assessed using their academic and personal data collected from 3 different colleges from Assam, India using two machine learning algorithms – Decision Tree, Naives Bayes. Here the accuracies are 60.29% and 73.53% respectively. These accuracies are optimized using three ensemble methods – Bagging, Boosting and Voting. The experimental results are reveals that, for Bagging, the accuracies are 69.21% and 75% for J48 and Naïve Bayes respectively. Then for Boosting, they are 70.59% and 72.06%. Then the two algorithms altogether used in Voting and the resultant accuracy is 63.24. This comparative study concludes that the three ensemble methods – Bagging, Boosting and Voting are well suitable for the Decision Tree (J48) as they improve the accuracy. But for Naïve Bayes, only Bagging proves to contain the best result. This study can be further extended in future with the other ensemble method named Stacking.

**REFERENCES**

[1] Romero, C., Ventura, S. (2007). Educational Data Mining: A Survey from 1995 to 2005. Expert Systems with Applications 33, 2007, pp.135-146.

[2] Sadiq Hussain, Neama Abdulaziz Dahan, Fadl Mutaher Ba-Alwi, Najoua Ribata (2018). Educational Data Mining and Analysis of Students' Academic Performance Using WEKA, Indonesian Journal of Electrical Engineering and Computer Science Vol. 9, No. 2, February 2018, pp. 447~459.

[3] Hijazi, S.T. and S.M.M.R. Naqvi, Factors Affecting Students' Performance, A Case Of Private Colleges. Bangladesh e-Journal of Sociology, 2006. 3(1): p. 10. 9.

[4] Bhardwaj, B.K. and S. Pal, Data Mining: A prediction for performance improvement using classification. International Journal of Computer Science and Information Security, 2011. 9(4): p. 5. 10.

[5] Strecht, P., et al., A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance. Proceedings of the 8th International Conference on Educational Data Mining, 2015: p. 3.

[6] Dekker, G.W., M. Pechenizkiy, and J.M. Vleeshouwers, Predicting students drop out: A case study. EDM '09Educational Data Mining 2009: 2nd International Conference on Educational Data Mining, 2009. 2: p. 10.

[7] Superby, J. Vandamme, J., Meskens, N. (2006). Determination of factors influencing the achievement of the first-year university students using data mining methods. Proceedings of the Workshop on Educational Data Mining at the 8th International Conference on Intelligent Tutoring Systems (ITS 2006). Jhongli, Taiwan, pp37-44.

[8] Vandamme, J., Meskens, N., Superby, J. (2007). Predicting Academic Performance by Data Mining Methods. Education Economics, 15(4), pp405-419.

[9] Cortez, P., Silva, A. (2008). Using Data Mining to Predict Secondary School Student Performance. EUROSIS, A. Brito and J. Teixeira (Eds.), 2008, pp.5-12.

[10] Noel-Levitz White Paper (2008). Qualifying Enrollment Success: Maximizing Student Recruitment and Retention Through Predictive Modeling. Noel-Levitz, Inc., 2008.

[11] Ajay Kumar Pal and Saurabh Pal, "Classification Model of Prediction for Placement of Students", I. J. Modern Education and Computer Science, 2013, 11, 49-56.

[12] A.G. Karegowdal, A. S. Manjunath2 and M.A. Jayaram3, "Comparative study of attribute selection using gain ratio and correlation based feature selection", International Journal of Information Technology and Knowledge Management, vol2, no.2, (2010),pp. 271 – 277.

[13] Z.H. Zhou, "Ensemble methods: Foundation and algorithms", CRC Press (2012).