INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 6.078
(Volume 6, Issue 5)
Available online at: www.ijariit.com

Using ensemble random forest, boosting and base classifiers to ameliorate prediction of students' academic performance

Olukoya Bamidele Musiliu
deleolukoya101@gmail.com
Federal University of Oye-Ekiti, Nigeria

ABSTRACT

In recent time, educational data mining (EDM) has received substantial considerations. Many techniques of data mining have been proposed to dig out out-of-sight knowledge in educational data. The Knowledge obtained assists the academic institutions to further enhance their process of learning and methods of passing knowledge to students. Education Data Mining have been playing substantial role in predicting student's academic performance. In this study, a novel student's performance prediction model premised on techniques of data mining with Students' Essential Features (SEF). Students' Essential Features (SEF) are linked to the learner's interactivity with the e-learning management system. The performance of student's predictive model is assessed by set of classifiers, viz. Bayes Network, Logistic Regression and REP Tree. Consequently, ensemble methods of Boosting and Random Forest using WEKA as an Open Source Tool are applied to improve the performance of these single classifiers. The results obtained reveal that there is a robust affinity between learner's behaviors and their academic attainment. Results from the study shows that REP Tree and its ensemble record the highest accuracy of 83.33% using SEF. Hence, in terms of Receiver Operating Curve (ROC), boosting method of REP Tree records 0.903, which is the best. This result further demonstrates the dependability of the proposed model.

Keywords: *EDM, Ensemble, Boosting, Random Forest, Data mining, Classifiers, machine learning*

1. INTRODUCTION

From the last decade there has been the advent of information technology in several fields and this has led to the availability of large volumes of data repository in different formats such as; files, documents, sound, records, scientific, video data and many other data formats. This data collection does not exclude the academic environment, all these data collected from diverse platform demand for proper method of deducing facts or knowledge from the big data for better decision making. Knowledge Discovery in Databases (KDD), is also known as data mining, which purposes is the discovery of useful information from big data collected or gathered (Brijesh & Saurabh, 1996). The main functions of data mining are applying various methods and algorithms in order to discover and extract patterns of stored data (Nikhil & Rudresh, 2012). In the environment of educational data mining (EDM), several data mining techniques have been exploited by researchers with the purpose of discovering hidden patterns from educational settings. But, due to the eruption of self-governing academic data sources, there is a growing demand for application of effective data mining techniques in this direction, so that performance of students can be improvised at large.

Educational data mining (EDM) is concerned with developing, researching, and applying computerized methods to detect patterns in large collections of educational data that would otherwise be hard or impossible to analyze due to the enormous volume of data within which they exist (Romero, 2010). EDM has emerged as a research area in recent years aimed at analyzing the unique kinds of data that arise in educational settings to resolve educational research issues (Baker & Yacef, 2009). When the concept of machine learning is applied in educational domain, then it is referred to as Educational Data Mining (EDM). In fact, EDM, can be defined as the application of data mining (DM) techniques to this specific type of dataset that come from educational environments to address important educational questions (Romero, 2010). To classify or predict, there are several classification learning schemes such as logistic regression (example of the algorithms is LogitBoost), decision tree (with algorithms like ID3, C4.5, REP Tree, CART etc.), backtracking (ANN such as MLP, etc.), probability (such as Bayes Network, Naïve Bayes etc.) and many other learning schemes that are highly embraced to mine educational data. When any of these algorithms is used to do classification as a lone classifier, it is referred to as a single classifier (Gupta et al., 2015). All these single classifiers are attached with one limitation or the others. Hence the need to improve their performance gives rise to ensemble method that combines many classifiers to form a unit. There are various ensemble methods such as Bagging, Boosting, Random forest, Stacking, Voting etc. (Shet, 2014).

Educational constituted authorities at all levels are exploring to integrate predictive components into their pedagogical environment to help students'. Furthermore, EDM and utilization of various tools have shown a significant improvement in the

advancement of student's carrier. Therefore, as a matter of consequence, the ability to foretell the performance of students becomes imperative in academic settings. Yet it was observed that there are still areas that can be further explored, other algorithms can be applied in the ensemble frameworks with the aim of possibly increasing the prediction accuracy. Thus, this study is posed to explore other algorithms of different learning schemes of decision tree, logistic regression, backtracking and probability two ensemble methods (Boosting and Random forest) to predict students' performance for comparison purpose.

2. LITERATURE REVIEW

Tama (2015) employed data mining to prevent the rate of inactive students. The study explored ensemble methods to unravel the major factors that influenced non-activeness of students in the school. The result from the study revealed that demographic attributes such as marital status and employment status in the society play a vital role in students' activeness. Also, it was shown in the study that rotation forest has the best prediction accuracy compared to other classifiers. The study conducted by Wati et al. (2017) was carried out to unveil the causes of poor learning attitude of students in the school. The study focused on comparison of the performances of two data mining algorithms to predict student learning based on the student records (data set). After the experiment, the result showed that average percentage of both classifiers was above 60%, whereas Naïve Bayes has higher precision average. Mhetre (2017) conducted a study to identify slow, average and fast learners among students. The study employed the techniques of data mining to predict the academic performance of student. Naive Bayes, J48, Zero and Random Tree models were trained and tested on the dataset. In the end, the result showed that Random Forest has higher accuracy over other three algorithms. The study by Amrieh et al. (2016) was conducted to predict the academic performance of newly admitted student. Three classification techniques of machine learning were proposed and implemented to model the new data attributes features obtained from learner's interactivity e-learning management system. Sequence to this, the result showed that learners' behavior has impact on student's academic performance. Baradwaj & Pal (2012) studied the use of classification data mining task to evaluate the academic performance of student. ID3 was applied to model student's information. The study submitted that factors like attendance, class test, seminar and assignment etc. contribute to the dropout of students from their various studies. Since ID3 algorithm cannot handle numerical and missing values, therefore the study could have considered other decision algorithms like C4.5. The case study reported by Thakar et al. (2015) surveyed the Educational Data Mining (EDM) and its scope. After series of investigation, it was submitted that most of the studies are segregated and there is need for integrated methods. The study projected many factors that influence the performance of students in academic environment. Arora, Singhal, & Bansal (2014) conducted a study to improve the quality of education by using the available parameters that is relevant to students' performance in citadel of learning. The study proposed Radial Basis Function (RBF) using Neural Network to predict marks obtained from courses. The results from the study showed that the proposed Radial Basis Function (RBF) using Neural Network is an effective tool to avert mishaps in students' performance.

2.1 Concept of Data Mining

Data mining is the process of analyzing large amount of data from repository to unveil hidden patterns (knowledge) that are currently unknown and that are possibly useful in supporting the decision-making process (Rokach, 2005). Generally, data mining is carried out with the aid of using mathematical, statistical, artificial intelligence, and machine learning techniques. Data mining could therefore be defined as application of machine learning techniques on data set, based on the commonest technique often deployed.

Data mining is also termed as Knowledge Discovery in Database (KDD), it simply means extraction or "mining" information from large repository of data. Data mining techniques are used to explore large volumes of data with a view to discovering hidden patterns for decision making process. While data mining and knowledge discovery in database are mostly used interchangeably, data mining is actually part of the knowledge discovery process (Baradwaj & Pal, 2012). Data mining refers to a particular step in the Knowledge discovery process. It consists of algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns (models) over the data (Rokach, 2005). Data mining is also a process of extracting nontrivial, valid, novel and useful information from large databases. Data mining can be viewed as a kind of search for meaningful patterns or rules from a large search space, that is, the database (Thakar et al., 2015).

2.3 Data Mining Tasks

The patterns to be discovered depend on the data mining tasks applied (Badr, Din, & Elaraby, 2014). Broadly categorized, data mining tasks are as follows:

a. Predictive tasks: these tasks use the output to characterize the general properties of data. These include: Clustering, Summarization, Association, Sequence Analysis, etc.

b. Descriptive tasks: these tasks perform functionalities on the provided dataset to predict how similar data set will behave. These include: Classification, Predication, Time series analysis, Regression etc.

2.4 Learning Scheme

This refers to the technique deployed in the construction of models i.e. how the models are trained.
There are following learning scheme in data mining but not limited to:
(1) **Decision tree:** example of algorithms are: ID3, C4.5, REP Tree, CART etc.
(2) **Backtracking:** such as Neural Network (this could be Single Layer Perceptron, Multilayer Perceptron etc.)
(3) **Probability:** examples are Naïve Bayes, Bayes Belief Network etc.
(4) **Logistic Regression**: example of the algorithm is Logist Boost algorithm

2.5 Ensemble Methods—Increasing the Accuracy

Composite methods(ensemble methods) are reported as means for improving classifier and predictor accuracy (Moniz & Branco, 2017). *Bagging* and *boosting* are two such means that use a blend of models. Each joins a series of *k* (k is an integer) learned

models (classifiers or predictors), *M*1, *M*2 … *Mk*, with the aim of creating an improved composite model, *M*. Both bagging and boosting can be used for classification as well as prediction. Other methods include but not limited to Stacking, Voting and Random Forest. Since every model is characterized with various drawbacks, the ultimate objective then is to join the strength of all models in other to enhance the accuracy. Also, ensemble methods are more vigorous in the presence of noise.

2.6 Random Forest Ensemble Method

This is an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. Random Forest works like bagging but in an enhanced way. In addition to random sampling of data points as obtained in bagging, Random Forest also performs random sampling on features thereby obviating feature engineering. This randomness can be fused in many ways as highlighted below.

a) Arbitrarily choose F input features to split at each node (Forest-RI).
b) Build linear amalgamations of the input features to split at each node (Forest-RC).
c) Arbitrarily choose one of the F best splits at each node.

The algorithm for Random Forest is stated as follows.
Construction of each tree for classification task is done using the following algorithm:

i. Let the number of training cases be N, and the number of variables in the classifier be M.
ii. It is known that the number m of input variables to be deployed to find the decision at a node of the tree; m should be much less than M.
iii. Select a training set for this tree by selecting n times with replacement from all N obtainable training cases (i.e. take a bootstrap sample). Use the rest of the cases to calculate the error of the tree, by predicting their classes.
iv. For each node of the tree, randomly choose m variables on which to base the decision at that node. Compute the best split according to these m variables in the sampling tuples.
v. Each tree is fully grown and not pruned (as may be done in constructing a normal tree classifier).

For prediction or classification, a new sample is pushed down the tree. It is assigned the label of the training sample in the terminal node it ends up in. This procedure is iterated over all trees in the ensemble, and the average vote of all trees is reported as random forest prediction.

2.7 Boosting Ensemble Method

Boosting is a dual-step approach, where firstly, subsets of the original data are used to produce a series of averagely performing models and then secondly boost their performance by joining them together using a particular cost function (majority vote). Unlike bagging, in the classical boosting the subset creation is not random and depends upon the performance of the previous models. Every new subset contains the elements that were (likely to be) misclassified by previous models. In this method, the models are built sequentially since the knowledge of the previously misclassified tuples are used in the subsequent models. Hence, this method reduces bias but may be susceptible to over-fitting. The algorithm for AdaBoost is expressed below.

A boosting algorithm

Input:
E, a set of *d* data points with class labels;
j, the number of iterations (one classifier is generated per iteration);
a classification learning system

Output: A combining model.

Method:
(i) initialize the weight of each tuple in ***E*** to 1=*d*;
(ii) for ***i*** = 1 to ***j*** do
(iii) sample ***E*** with replacement giving credence to the tuple weights to obtain ***D****i*;
(iv) utilize the training set ***E****i* to generate a model, ***M****i*;
(v) find *error*(***M****i*), the error rate of *Mi* (Equation 2.0)
(vi) if *error*(***M****i*) is more than ½ then
(vii) reinitialize the weights to 1=*d*
(viii) go back to step (iii) and try again;
(ix) endif
(x) for each tuple in ***E****i* that was accurately classified do
(xi) multiply the weight of the tuple by *equation 2.0*
(xii) normalize the weight of each tuple;
(xiii) endfor

Application of the combining model to classify tuple, ***Z***:
(i) initialize weight of each class to zero;
(ii) for *i* = 1 to *k* do:
(iii) ***W****i* = *equation 2.1;*

(iv) $f = \boldsymbol{Mi}(\boldsymbol{Z})$;
(v) add $\boldsymbol{Wi}$ to weight for class c
(vi) endfor
(vii) return the class with the largest weight;

Computation of the error rate of model *Mi* is done as follow.
We take the aggregate sum of the weights of each of the tuples in *Ei* that *Mi* misclassified. Mathematically,

$$error(Mi) = \sum_j^d wj \times err(Zj) \qquad 1$$

It is a good omen for classifier's error rate to go lesser as this result to more accurate the classifier becomes. Consequently, the higher its weight for voting would be. The weight of classifier *Mi*'s vote is computed as follow

$$\log \frac{1-error(Mi)}{error(Mi)} \qquad 2$$

For each class, f, the aggregate sum of weights of each classifier that assigned class *f* to $\boldsymbol{Z}$ is taken. The champion is the class with uppermost sum and is returned as the class prediction for tuple $\boldsymbol{Z}$.

It will be observed that the initial base classifier is accomplished using weighting constants that are all the same. The weighting constants are enlarged for sampling tuples that are wrongly classified and shrank for sampling tuples that are perfectly classified in the subsequent iterations.

3. METHOD

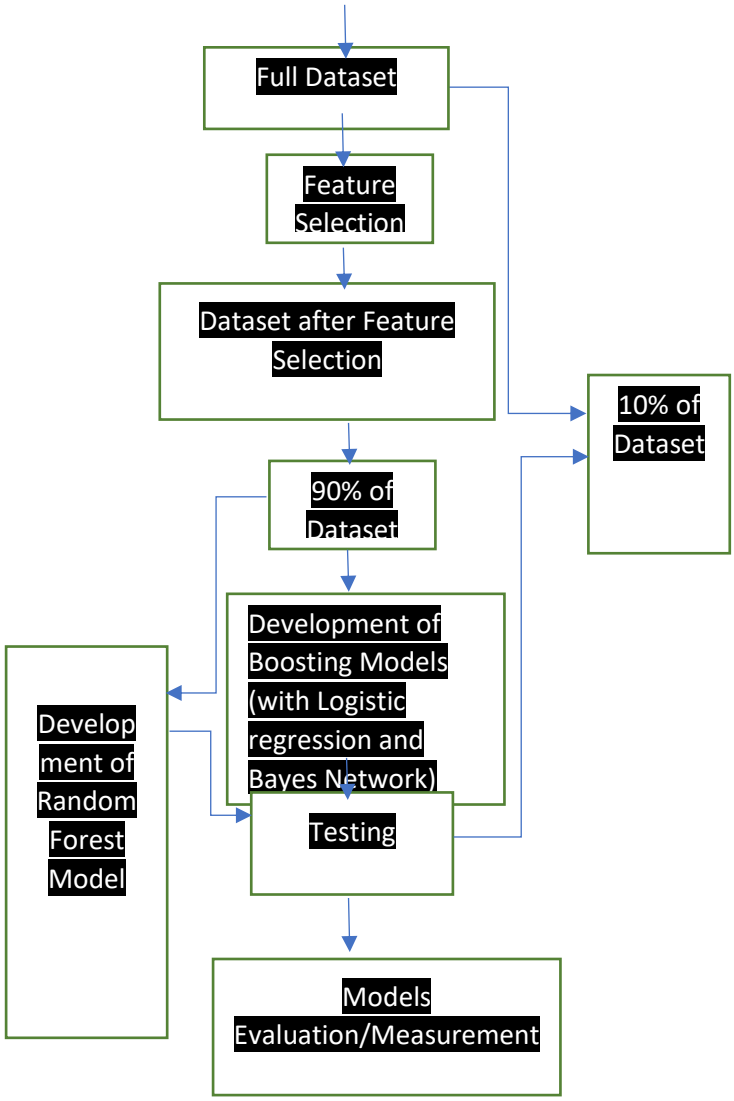

Fig. 1: System Model

3.1 Data Acquisition

The dataset used in this study was retrieved from Kaggle online machine learning repository and it is readily available for data mining. This dataset was originally collected from students through learning management system (LMS) called Kalboard 360. The dataset contains records of 480 students with 16 attributes. These attributes are classified into three categories as follows. Category 1 (Demographical Attributes): gender and nationality, place of birth and parent responsible for student. Category 2

(Education Background): educational stage, grade level and section. Category 3 (Behavioral Attributes): raised hand in class, visited resources, parent answering survey and parent school satisfaction.

3.2 Development of Ensemble Models

The system used in the development of ensemble models in this research is the Waikato Environment for Knowledge Analysis (WEKA). WEKA is a machine learning system developed by the University of Waikato in New Zealand that implements data mining algorithms using the JAVA programming language. WEKA is a contemporary tool for developing machine learning techniques and their deployment to actual domains of interest like crime detection and so on.

3.3 Effective Features

Correlation Based Feature Selection (CFS) is applied to decide the most important attributes in predicting student's performance. Table 1.0 presents the results of selected features. Search method of best first was used. CFS started set was with no attributes. Forward Search direction was employed in the search. Stale search after 5 node expansions was recorded. Total number of subsets evaluated stood at 121. Merit of best subset found set at 0.564.

Table 1: Result of CFS feature selection on the dataset

S no.	Selected Attributes
1	Relation
2	Raisedhands
3	VisITedResources
4	Discussion
5	ParentAnsweringSurvey
6	StudentAbsenceDays

As shown in the table above, 6 attributes are selected as most important out of total 16 attributes. Hence, these attributes returned by this feature selection method is referred to as Students' essential features (SEF) in this study.

3.4 Evaluation and Measurement Terms

The evaluation of the models is done using the confusion matrix values as basic measurement parameters. Its derivatives are also used.

Table 2: Confusion Matrix

		Calculated	
	+	**Positive**	**Negative**
Real	-	True Positive (T_p)	False Negative (F_N)
		False Positive (F_p)	True Negative (T_N)

Accuracy is the proportion of the total number of predictions where correctly calculated. Precision is the ratio of the correctly classified cases to the total number of misclassified cases and correctly classified cases. Recall is the ratio of correctly classified cases to the total number of unclassified cases and correctly classified cases. In addition, F-measure combine the recall and precision which is considered a good indicator of the relationship between them, while ROC Area, is a curve generated by plotting the true positive rate against the false positive rate at various threshold parameter adjustments. Measurement of accuracy is done using the area under the ROC curve. A perfect model should have its ROC Area = 1, the closer the ROC value to 0.5 the worse it is.

$$\boldsymbol{Accuracy} = \frac{T_P + T_N}{T_P + F_N + F_P + T_N}$$

$$\boldsymbol{Precion} = \frac{T_P}{T_P + F_P}$$

$$\boldsymbol{Recall} = \frac{T_P}{T_P + F_N}$$

$$\boldsymbol{F.Measure} = \frac{Precion_c \times Recall_c}{Precion_c + Recall_c}$$

4. RESULTS ANALYSIS

Table 3: Results as Single Classifiers and as Ensemble Method during Training

Evaluation Measurement	Single classifiers			Boosting (Ensemble)		
Classifiers	**B N**	**L R**	**REP Tree**	**B N**	**L R**	**REP Tree**
Accuracy	74.7685	74.537	69.9074	74.7685	74.537	70.8333
Precision	0.749	0.745	0.698	0.749	0.745	0.707
Recall	0.748	0.745	0.699	0.748	0.745	0.708
F-Measure	0.747	0.745	0.698	0.747	0.745	0.707
ROC	0.868	0.871	0.822	0.845	0.812	0.844

Table 3 presents the result obtained during training as single classifiers and Boosting as ensemble method. This is not conclusive since it is a training stage, what matters most is the success records during testing stage. As observed in above, Bayes network perform better as single classifier in terms of accuracy with 74.7685% compare to other single classifiers and as boosting ensemble result. Also, it is observed that Logistic Regression returned better results of 0.871 of ROC as single classifier compared to others and ensemble method.

Table 4: Results as Single Classifiers and as Ensemble Method during Testing

Evaluation Measurement	Single classifiers			Boosting (Ensemble)		
Classifiers	**B N**	**L R**	**REP Tree**	**B N**	**L R**	**REP Tree**
Accuracy	81.25	81.25	83.3333	81.25	81.25	83.3333
Precision	0.742	0.811	0.827	0.742	0.810	0.827
Recall	0.813	0.813	0.833	0.813	0.792	0.833
F-Measure	0.775	0.808	0.823	0.775	0.783	0.823
ROC	0.921	0.961	0.851	0.897	0.927	0.903

The results obtained during the testing stage showed that there is no difference between the single classifier of Bayes Network, Logistic regression and RepTree and its experimented ensembles. Hence, boosting (Bayes Network) could be safely picked as being the best in this case. The results of Logistic Regression and its ensembles during testing stage. Just as observed in the case of Bayes Network also, the results showed that the prediction accuracy of students' performance are the same with single classifier of Logistic Regression and its ensembles. Logistic Regression and its ensembles recorded 81.25% prediction accuracy. However, choosing best classifier will involve the consideration of other factors such as Receiver Operating Curve (ROC). Single classifier of Logistic Regression records highest value for ROC with 0.961 which is the closest to 1. However, the results of prediction with REP Tree and Its Ensembles during testing stage. Same value is recorded across board here also. Single classifier of REP Tree and its ensembles returned prediction accuracy of 83.33 %. However, to select the best classifier will now involve the consideration of other factors such as Receiver Operating Curve (ROC). Boosting method records highest value for ROC with 0.903 which is the closest to 1.

Table 5: Performance Evaluation/Measurement of Random Forest

Parameters	Random Forest (Training)	Random Forest (Testing)
Correctly Classified Instances (%)	74.537	79.1667
Precision	0.746	0.806
Recall	0.745	0.792
F-Measure	0.745	0.784
ROC AREA	0.877	0.875

Table 5 presents the prediction results of Random Forest. It is observed that Random Forest could match not up with the previous classifier of Bayes Network, Logistic Regression and REP Tree, both as single classifiers and their ensembles.

With the results of this study, it can be deduced that the REP Tree performed better compared to others in predicting students' academic performance with prediction accuracy of 83.33%. This means that 30 of 38 students (testing set) are correctly classified to the right class labels (High, Medium and Low) and 8 students are incorrectly classified. The results of this study further prove the reliability of the proposed model. Compared to the study by Amrieh et al. (2016a), REP Tree performed better than all the single classifiers (C4.5, Neural Network and Naïve Bayes) deployed together with their ensemble of boosting and random forest in the study. The highest value recorded by Amrieh et al. (2016a) for prediction during testing is 79.1% while 82.2% is recorded during validation.

5. CONCLUSION

Consequently, ensemble methods are applied to improve the performance of these single classifiers. Bagging, Boosting and Random Forest (RF), which form the array of most frequently used ensemble methods as reported in different literature are deployed. The obtained results reveal that there is a strong relationship between student's essential features and their academic achievement. The accuracy of student's predictive model using students' essential features in the case of REP Tree as single classifier and in ensemble methods achieved 83.33% prediction accuracy. In terms of ROC, boosting method of REP Tree achieved best with 0.903.

6. REFERENCES

[1] Amrieh, E. A., Hamtini, T., and Aljarah, I. (2016). Mining Educational Data to Predict Student's Academic Performance Using Ensemble Methods. International Journal of Database Theory and Application, 9(8), 209–213. https://doi.org/10.14257/ijdta.2016.9.8.13

[2] Badr, A., Din, E., and Elaraby, I. S. (2014). Data Mining : A Prediction for Student â€™ s Performance Using Classification Method. World Journal of Computer Application and Technology, 2(2), 43–47. https://doi.org/10.13189/wjcat.2014.020203

[3] Baker, RSJd and Yacef K. (2009). The State of Educational Data Mining in: A Review and Future Visions. JEdu Data Min Pp.3–17.

[4] Baradwaj, B., and Pal, S. (2012). Mining Educational Data to Analyze Student's Performance. Internation Journal of Advamced Computer Science and Applications, 2(6), 63–69. https://doi.org/vol.2,No.6

[5] Brijesh, K. B and Saurabh, P (1996). Data mining: machine learning, statistics, and databases,.

[6] Gupta, A., Gupta, S., and Singh, D. (2015). A Systematic Review of Classification Techniques and Implementation of ID3 Decision Tree Algorithm. In International Conference on System Modeling & Advancement in Research Trends (SMART) (pp. 144–152).
[7] Mhetre, V. (2017). Classification Based Data Mining Algorithms to Predict Slow, Average and Fast Learners in Educational System Using Weka. In Proceedings of the IEEE 2017 International Conference on Computing Methodologies and Communication (pp. 475–479).
[8] Moniz, N., & Branco, P. (2017). Evaluation of Ensemble Methods in Imbalanced Regression Tasks. In Proceedings of Machine Learning Research (pp. 129–140).
[9] Nikhil, R. and Rudresh, S (2012). Data Mining on Educational Domain.
[10] Rokach, L. (2005). Ensemble Methods for Classification. In Data Mining and Knowledge Discovery Handbook (pp. 957–980). https://doi.org/10.1007/0-387-25465-X_45
[11] Romero, C, and Ventura, S. (2010). Educational data mining: a review of the state-of-the-art. IEEE Trans Syst Man Cybern C: Appl Rev 2010, 40:601–618.
[12] Romero, C., Ventura, S., Pechenizky, M., and Baker, R. (2010). Hand-book of Educational Data Mining. Data Mining and Knowledge Discovery Series. Boca Raton, FL: Chapman and Hall/CRC Press
[13] Shet, S. (2014). Approach for Predicting Student Performance Using Ensemble Model Method. International Journal of Innovative Research in Computer and Communication Engineering, 2(5), 161–169.
[14] Tama, B. A. (2015). Learning to Prevent Inactive Student of Indonesia Open University. Journal of Information Processing Systems (JIPS), 11(2), 165–172.
[15] Thakar, P., Mehta, A., and Manisha. (2015). Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue. International Journal of Computer Applications, 110(15), 60–68.
[16] Wati, M., Indrawan, W., Widians, J. A., and Puspitasari, N. (2017). Data Mining For Predicting Students ' Learning Result. In Dept. of Computer Science and Information Technology Universitas Mulawarman Samarinda, Indonesia (p. 28).

