Bayero Journal of Pure and Applied Sciences, 11(2): 142 - 148
Received: November, 2018
Accepted: December, 2018
ISSN 2006 – 6996

APPLICATION OF CLASSIFICATION MODELS TO PREDICT STUDENTS' ACADEMIC PERFORMANCE USING CLASSIFIERS ENSEMBLE AND SYNTHETIC MINORITY OVER SAMPLING TECHNIQUES

[1]Abdulazeez, Y and [2]Abdulwahab, L.
[1]Department of Computer Science, Federal University, Duste, Jigawa State
[2]Department of Information Technology, Bayero University, Kano
Corresponding author: abd_wahhb@yahoo.com

ABSTRACT
The demand for data-driven decision making has resulted in the application of data mining in the educational sector and other disciplines. The needs for improving the performance of data mining models have been identified as an interesting area of research globally. Higher educational institutions keep a large amount of students' data, but these data are rarely used effectively in decision and or policy-making processes. This research is an attempt to enhance the performance of data mining models to predict students' academic performance using stacking classifiers ensemble and synthetic minority over-sampling techniques. The three (3) classifiers models J48, IBK and SMO were trained and tested on 206 students' data set using previous academic performance records of Federal University Dutse, Nigeria. WEKA 3.9.1 data mining tool was used in predicting the final year student's classes of degree at an undergraduate level, while Unified Tertiary Matriculation Examination, Senior Secondary Certificate Examinations and first-year Cumulative Grade Point Average of students served as inputs to the model. The result obtained showed that on training dataset after class balancing, stacking classifiers ensemble model out-performing the other three (3) classifiers models in both performance accuracy (96.7949%) and RSME (0.1098), suggesting that stacking classifiers ensemble is the best model in context of this research.
Keywords: Educational Data Mining, J48. SMO. IBK, Stacking Classifiers Ensemble

INTRODUCTION

Decision making is gradually becoming data-driven with advancement in information and communication technology (ICT). This is possible due to a large amount of data generated. Data mining has a wide range of application in various disciplines (Baker & Yacef, 2009). Recently, Data mining is widely used on educational dataset often referred to as educational data mining (EDM). Kaur, Manpreet and Gurpreet (2015) opined that EDM is concerned with developing methods that discover knowledge in data originating from educational environments, using different data mining techniques and machine learning algorithms. Natek and Zwillingi (2014) suggested that some of the problems relating to students' success in a course are hard to solve simply because normal statistical methods are not significant enough to uncover the hidden patterns and knowledge, useful for educational processes planning and organization.

Therefore, the needs to adopt a data mining technique for solving problems relating to students' success using data originating from educational environments becomes crucial.

Nithya Umamaheswari and Umadevi (2016) categorized various methods used in EDM into the following: Classification, Clustering, Relationship Mining, Discovery with Models and Distillation of data for human judgment. These data mining methods have been applied in prior research and were reported to have promising performance (Nithya *et al.,* 2016).

Prior studies have been conducted on predicting students' academic performance using various data mining techniques and machine learning algorithms (Jadrić. Željko and Maja 2010; Jishan *et al.,* 2015; Kabra and Bichar, 2011). Nikolovski *et al.* (2015) adopted two classifiers algorithms in predicting the dropout features of students. Gray, Colm and Philip (2014) used additional classifier algorithms namely: Neural Network (NN), Decision Tree, Support Vector Machine (SVM), K-nearest neighbor (KNN), Naïve Bayes and Logistic Regression to predict learners' progression in tertiary education. The result obtained suggested SVM has the highest performance accuracy of 73.33% and the least performance was recorded by Logistic Regression which has 60.05% accuracy. Kaur *et al.* (2015) focused on identifying the slow learners among students and displaying it by a predictive data mining model using classification based algorithms Multilayer Perception, Naïve Bayes, Sequential Minimal Optimization (SMO), J48 and REPTree of all the five algorithms, multilayer Perception has the highest accuracy of 75%.

This research adopts Cross Industry Standard Process for Data Mining (CRISP-DM), using machine learning algorithms J48 decision tree classifier, SMO classifier and-Nearest Neighbors (IBK) classifier.

The objective of this research is to develop a stacking classifiers ensemble data mining model that predicts students academic performances based on the three classifiers J48, IBK and SMO (Sen *et al.,* 2012).

MATERIALS AND METHODS

The methodology for EDM is not yet clearly defined and there are no clear standards about which data mining methods or algorithms are preferable in this context. Various data mining methods have been used by different researchers for estimating preferable algorithms (Mythili and Mohamed, 2014; Nikolovski *et al.,*, 2015). In general Data mining processes follows a set of steps that must be executed regardless of the algorithms or methodology that will be implemented (Tabra and Lawan, 2017).

Data Collection

A total of 206 students' data from the Faculty of Science, Federal University Dutse was collected. The data set was divided into two subsets for model training and testing respectively. The model was trained using 164 students data representing 80% of the data set while 42 (20%) students' data set was used in model testing. The dataset collected contains information like student registration number, students Name, Phone Number and Address which are not needed in conducting the research work as such the data set has to be balanced using SMOTE to avoid producing misleading results about the model performance. The students' attributes that are not needed in the data set collected for conducting the research has to be removed. To do this the data set has to undergo data preprocessing activities in order to prepare the data for our research.

Data Preparation and Cleaning

The data preparation phase covers all activities required in constructing the final data set that was fed into WEKA 3.9.1 data mining tools from the initial raw data. It is a known fact that real-world data tend to be incomplete, inconsistent and noisy. Therefore, for real-world data to be utilized by the data mining tool have to be further pre-processed. The manual method of data cleaning was applied to the collected data to remove noise and inconsistencies in the data. The students' data used for this research was pre-processed to ensure that information relating to the identity of the student is removed, and other irrelevant attributes are also removed using the attribute filter of WEKA 3.9.1. For the purpose of this research, only information relating to the previous academic performance of students were retained.

Descriptions of the Pre-Processed Data Set

The data was pre-processed to make it suitable for data mining. WEKA 3.9.1 data mining tool has the ability to pre-process data and remove irrelevant attributes from the dataset. The original data collection contains 20 attributes but some students' attributes in the original data set are not needed in conducting the research and were consequently expunge. A brief description of the original data collected is presented in Table 1. After data pre-processing ten (10) students attributes relating to previous academic performance records of students from the initial 20 were selected as relevant students' attributes for predicting students future academic performance. This is described in Table 2. Class imbalance problem was treated using SMOTE in WEKA 3.9.1 data mining tool. SMOTE allows for the simulation of instances in some class to make all the classes in the dataset equal.

Table 1. Attributes and Data Type of the Original Dataset

S/N	ATTRIBUTES	DATA TYPE
1	Registration Number	String
2	Candidate Name	String
3	State Of Origin	String
4	Local Government	String
5	Sex	String
6	Age	Integer
7	English Score	Float
8	Subject 2	String
9	Subject2 Score	Float
10	Subject 3	String
11	Subject3 Score	Float
12	Subject 4	String
13	Subject 4 Score	Float
14	Total Score	Float
15	English Grade	String
16	Subject2 Grade	String
17	Subject3 Grade	String
18	Subject4 Grade	String
19	First CGPA	Float
20	Predicted Class of Graduation	String

The original data set contains 20 attributes, some attributes are not needed in developing the model and as such, they were filtered and removed during pre-processing as shown in Table 2.

Table 2: Summary of Selected Students 'Attributes

S/N	ATTRIBUTES	DATA TYPE
1	English Score	Float
2	Subject2 Score	Float
3	Subject3 Score	Float
4	Subject 4 Score	Float
5	English Grade	String
6	Subject2 Grade	String
7	Subject 3 Grade	String
8	Subject 4 Grade	String
9	First Year CGPA	Float
10	Predicted Class of Graduation	Nominal

Modeling

The research work attempts to develop a stacking classifier ensemble data mining model. The machine learning algorithms adopted are J48 decision tree classifier, SMO classifier and- IBK classifier and stack ensemble. SMOTE was used for balancing the classes in the data set thus, increasing the volume of the training data set from 164 instances to 312 instances thereby making all the four classes to have 78 equal numbers of instances. Figures 1 and 2 showed Dataset before and after Class balancing with SMOTE on WEKA Explorer

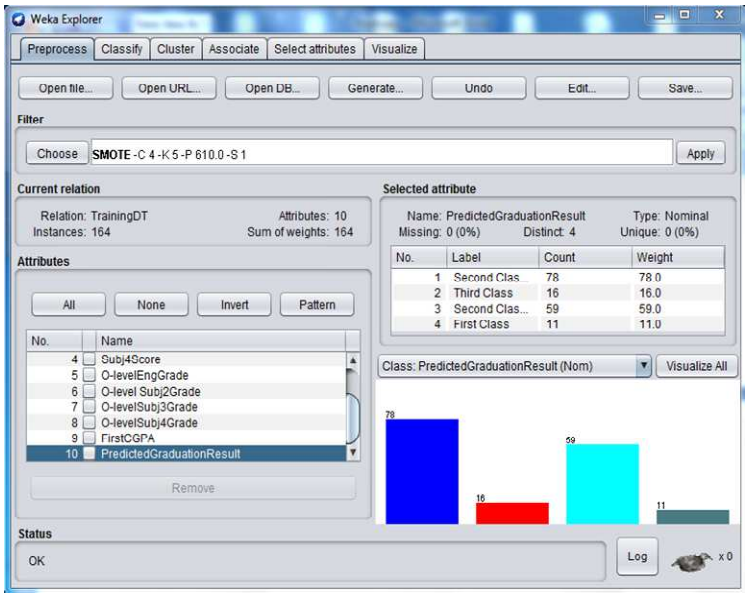

Figure 1. Dataset before Class Balancing

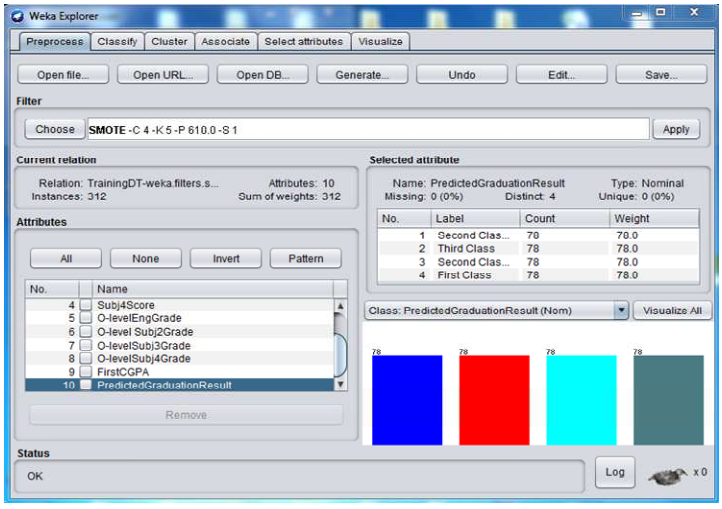

Figure 2. Dataset after Class Balancing with SMOTE on WEKA Explorer

The three (3) machine learning algorithms were trained and tested using 10-fold cross -validation to avoid over-fitting the models. Figure 3 depicts the proposed model using machine learning algorithms.

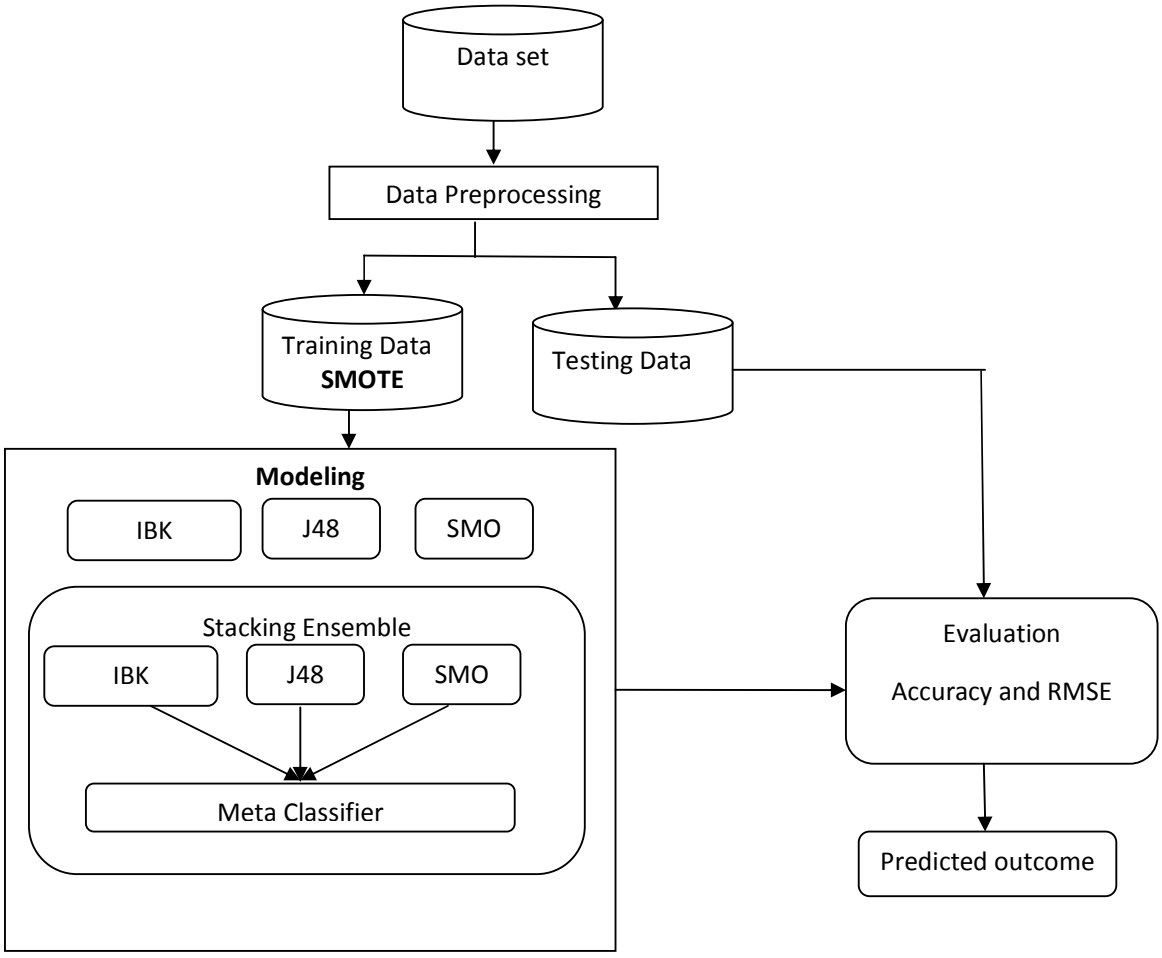

Figure 3. Proposed Model

Model Training and Testing

In this research, a series of training and testing were carried out using the algorithms. The data set was divided into two subsets for model training and testing. For training, 80 % of the data set was used and the remaining 20% of the data set was used for testing. Ten (10)-fold cross validation was used throughout model training and testing to avoid over fitting the models. Since the data set is small and imbalanced SMOTE technique was used to balance the classes and increase data volume of the training data sets. The WEKA 3.9.1 data mining tool provides a training and testing option to train and test on the same data set, the classifiers were trained on (training set), and also tested on a user-specified test data (supplied test set).

Performance accuracies of the various models were recorded during training and testing. Before feeding the data set to WEKA 3.9.1 data mining tool for testing, the actual performance of the students in the last column "Predicted Graduation Result" was left blank. The test was conducted using the same method for all the models.

Model Evaluation

To evaluate the performance of the various models the confusion matrices of the models were empirically evaluated to select among the different models (Duda, Peter & David, 2000). Percentage of models' performance accuracy should not be the only metric to be used for evaluating data mining model performance. Other performance metrics should also be considered. Performance accuracy and Root mean square error (RMSE) was used to indicate the various model performances and also define RMSE as a measure of the differences between values predicted by the model and the values actually observed (Nipaporn *et al.,* 2016). If a value of RMSE is high, it means that the predicted values are scattered away from the average predicted value. If their values are low, then the predicted value tend to cluster close to the predicted average. This is calculated using the formula in equation 1.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(\hat{Y}_i - Y_i\right)^2}{N}} \quad (1)$$

Where $\hat{Y}_i$ is the predicted data generated from the model, *Yi* is the actual value, and *N* is the total number of data. Detail discussion of binary classification was presented because other evaluation measures were derived from the binary confusion matrix and its performance measures. Table 3 shows a general binary confusion matrix template.

Table 3 Binary Confusion Matrix Template

		Predicted Performance		TOTAL
		NEGATIVE	POSITIVE	
Actual Performance	NEGATIVE	TN	FP	TN+FP
	POSITIVE	FN	TP	TP+FN
TOTAL		TN+FN	FP+TP	TN+FP+FN+TP

Key: TN=True Negative FP=False Positive FN=False Negative TP=True Positive

True Positive Rate (TPR) is the number of correctly classified instances in a given class. It is also known as sensitivity or hit ratio and can be defined as

$$\text{TPR} = \frac{TP}{\text{TP+FN}} \quad (2)$$

False Positive Rate (FPR) is the number of incorrectly classified instances of a given class. It is also known as fall out and can be defined as

$$FPR = \frac{\text{FP}}{\text{FP+TN}} \quad (3)$$

The accuracy of the classifier is the total number of correctly classified instances. It can be defined as

$$Accuracy = \frac{TP+TN}{\text{TP+FN+ FP+TN}} \quad (4)$$

RESULTS AND DISCUSSION

The results of training the various models using the training data set before class balancing are presented in Table 4. The results of training the various models' using the training data set after class balancing are presented in Table 5. The result obtained suggested that class balancing using SMOTE results in improving all the various models performance. The stacking classifiers ensemble model outperformed the other models in both performance accuracy and RSME values which makes the model better than the other classifiers model.

Table 4: Performance Result of Various Classifiers on Training Dataset Before Class Balancing

S/N	Classifiers	Accuracy %	RMSE	TPR	FPR	Precision
1	J48	87.1951	0.2322	0.872	0.082	0.881
2	SMO	86.5854	0.3301	0.866	0.096	0.870
3	IBK	82.9268	0.2097	0.829	0.134	0.840
4	Stacking Ensemble	87.1951	0.2404	0.872	0.082	0.881

Table 5: Performance Result of Various Classifiers on Training Dataset after Class Balancing

S/N	Classifiers	Accuracy %	RMSE	TPR	FPR	Precision
1	J48	95.1923	0.1455	0.952	0.016	0.953
2	SMO	90.7051	0.3248	0.907	0.031	0.908
3	IBK	90.7051	0.1591	0.907	0.031	0.919
4	Stacking Ensemble	96.7949	0.1098	0.968	0.011	0.969

The various models' performance accuracy results obtained on testing the individual classifier models on the test dataset indicated that the stacking ensemble model outperformed the other three (3) models with an accuracy of 96. 8%. The Models Performance Accuracy based on Testing Data is shown in Figure 4.

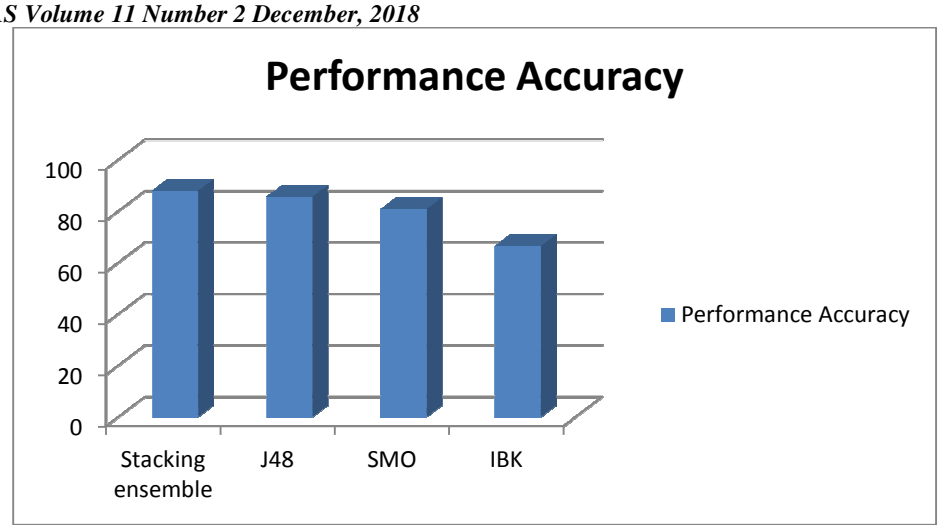

Figure 4: Bar chart of the Models Performance Accuracy on Testing Data

Predicting the academic performance of students is a challenging task vis a vis that students' academic performance depend on diverse factors such as personal, socio-economic, psychological and other environmental factors. Satyanarayana and Nuckowski (2016) identified ensemble methods as the most influential development in Data Mining and Machine Learning. Classifier ensembles include a combination of the multiple models into one usually more accurate than the best of its components. An approach for predicting students' academic performance using the ensemble model method was presented in the prior research (Shet & Gayathri, 2014). Stacking ensemble technique was used in predicting the academic achievement of students (Nippon *et al.,* 2016). Performance of the three classifiers algorithms was evaluated and the stacking ensemble technique was used to combine the three classifiers. Stacking ensemble techniques used in this research aimed at improving models performance. This technique combines various classifiers output as an input in a Meta classifier. The standard stacking technique is presented in Figure 5. Stacking ensemble technique has the capability of combining heterogeneous base classifiers and a Meta classifier is trained for final prediction (Dzeroski & Bernard, 2004). Predicting the output of base classifiers are fed directly as data input into the Meta classifier for training and final prediction.

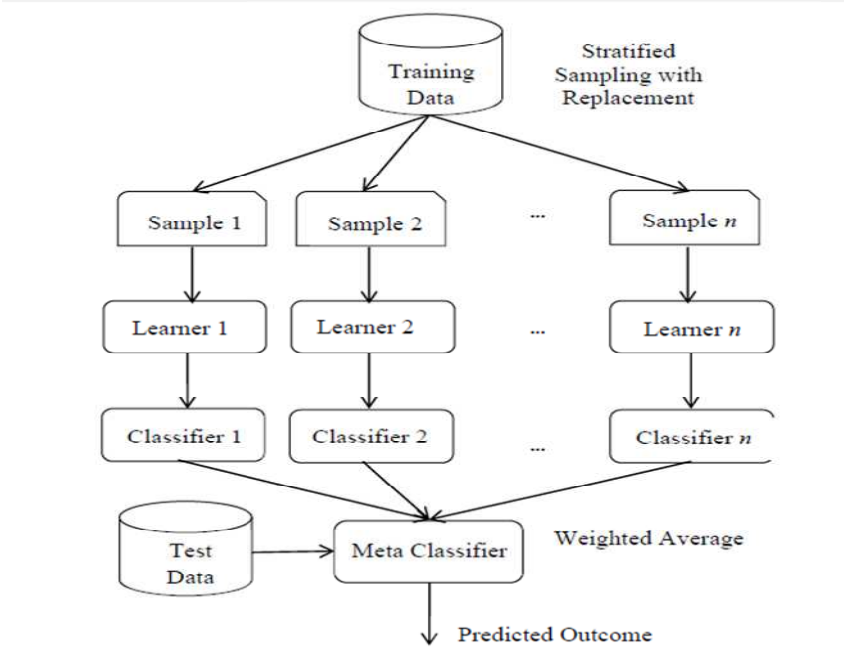

Figure 5. Standard Stacking Ensemble Technique (Source: Sikora and Al-laymoun, 2014)

CONCLUSION

Stacking classifiers ensemble techniques has the capability of improving the performance accuracy and efficiency of students' academic performance prediction models. The findings from this research showed that the stacking classifiers ensemble model performance accuracy and RMSE values are better than the other three individual classifier models suggesting that stacking classifiers ensemble is the best model in the context of this research. Though numerous machine learning algorithms exists but this research evaluated the performances of only three (3) Machine learning algorithms in developing the student academic performance prediction model. Further research can be conducted using other algorithms and consider other factors that could likely affect the academic performance of students such as socio-economic factors and psychological factors.

Contributions of Authors

Both authors wrote, conceived designed and performed the data mining analysis

Conflict of Interest

Authors have declared that no conflicting Interest exist

Acknowledgement

The authors wish to acknowledge the support from Bayero University, Kano for sponsoring this research..

REFERENCES

Baker, R. and Yacef K. (2009) "The State of Educational Data Mining in 2009: A Review and Future Visions" *Journal of Educational Data Mining, Article 1, Vol. 1, No.1 2009*

Duda R.O., Hart, P. E and Stock, D. G (2000). " Pattern Classification" Second Edition, Chapter Nine pp.532 - 536 John Willey & Sons.

Dzeroski, S and Bernard Z. (2004) "Is Combining Classifiers with Stacking Better than Selecting the Best One?" *Springer Journal of Machine Learning, Volume 54, pp. 255–273, 2004*

Gray, G, Colm M and Philip O. (2014) "An Application of Classification Models to Predict Learner Progression in Tertiary Education" *Conference Paper · February 2014 DOI: 10.1109/IAdCC.2014.6779384*

Jadrić, M. Željko G and Maja, Ć,(2010) "Student Dropout Analysis with Application of Data Mining Methods*"* Management, Vol. 15, 2010, 1, Pp. 31-46

Jishan, S.T, Raisul I.R Naheena H and Rashedur R. (2015)"Improving Accuracy of Students' Final Grade Prediction Model Using Optimal Equal Width Binning and Synthetic Minority Over-Sampling Technique" *Decision Analytics (2015) 2:1 A Springer Open Journal*

Kabra, R.R., Bichkar R.S, (2011) "Performance Prediction of Engineering Students Using

Decision trees", *International Journal of computer applications (0975-8887), Vol-36-No.11, December 2011.*

Kaur, P, Manpreet S and Gurpreet S. J (2015) **"**Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector" *3rd International Conference on Recent Trends in Computing 2015(ICRTC-2015) Procedia Computer Science 57 (2015) pp. 500 – 508doi: 10.1016/j.procs.2015.07.372*

Mythili, M.S. and Mohamed S.A.R., (2014) "An Analysis of students' Performance Using Classification Algorithms" *IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 1, Ver. III (Jan. 2014), pp. 63-69 www.iosrjournals.org*

Natek, S and Zwilling M. (2014) "Student Data Mining Solution–Knowledge Management System Related to Higher Education Institutions" *Expert Systems with Applications 41 (2014) pp. 6400–6407.Contents lists available at ScienceDirect*

Nikolovski. V, Stojanov R, Igor M, Ivan C, Gjorgji M. (2015) "Educational Data Mining: Case Study for Predicting Student Dropout in Higher Education" https://www.researchgate.net/publication/282333827 Conference Paper · April 2015

Nipaporn, C, Kreangsak T and Punnee S. (2016) "Stacking Technique for Academic Achievement Prediction" *2016 International Workshop on Smart Info-Media Systems in Asia (SISA 2016) pp.14-17*

Nithya. P., Umamaheswari B and Umadevi A. (2016) "*A Survey on Educational Data Mining inField of Education*" *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 5 Issue 1, pp.69 -78 January 2016*

Osmanbegović, E. and Mirza S. (2012) "Data Mining Approach For Predicting Student Performance" Economic Review – *Journal of Economics and Business, Vol. X, Issue 1, May 2012 pp. 4 -12.*

Pandey, M and Taruna S. (2014) "A Multi-Level Classification Model Pertaining to the Student's Academic Performance Prediction*"* *International Journal of Advances in Engineering & Technology, September, 2014. 1329 Vol. 7, Issue 4, pp. 1329-1341 ©IJAET ISSN: 22311963*

Satyanarayana, A and Nuckowski M. (2016) "Data Mining using EnsembleClassifiers for Improved Prediction of Student Academic Performance" *In ASEE Mid-Atlantic Section Spring 2016 Conference, George Washington University, Washington D.C*, April 8-9, 2016.

Sen, B., Emine, U and Dursun D. (2012) "Predicting and Analyzing SecondaryEducation placement-test scores: A data mining approach" *Expert Systems with Applications 39 (2012) pp. 9468–9476*

Shet, S and Gayathri, J. (2014) "Approach for Predicting Student Performance Using Ensemble Model Method" *International Journal of Innovative Research in Computer and Communication EngineeringVol.2, Special Issue 5, October 2014*pp. 161 – 169

Sikora, R and Al-Laymoun, O.H. (2014)., "A Modified Stacking Ensemble Machine Learning Algorithm Using Genetic Algorithms", *Journal of International Technology and Information Management*. vol.23, No.1, pp.1-12.

Tabra, M.S. and Lawan A. (2017) "A Comparative Analysis of the Performance of Three

Machine Learning Algorithms for Tweets on Nigerian Dataset" The International Journal of E-learning and Educational Technologies in th Digital Media (IJEETDM) 3(1) pp. 23-30.

Tekin, A. (2014) "Early Prediction of Students' Grade Point Averages at Graduation: A Data Mining Approach" *Eurasian Journal of Educational Research, Issue 54, 2014, pp. 207-226*