International Journal of Informatics and Communication Technology (IJ-ICT)
Vol.8, No.3, December 2019, pp. 122~127
ISSN: 2252-8776, DOI: 10.11591/ijict.v8i3.pp122-127

Classifiers ensemble and synthetic minority oversampling techniques for academic performance prediction

Abdulazeez Yusuf[1], Ayuba John[2]
[1]Department of Computer Science, Federal University Dutse, Nigeria
[2]Department of Cyber Security, Federal University Dutse, Nigeria

Article Info

Article history:

Received Sep 18, 2019
Revised Nov 3, 2019
Accepted Nov 23, 2019

Keywords:

Academic performance prediction
Educational data mining (EDM)
Stacking classifiers ensemble
Synthetic minority over-sampling technique (SMOTE)

ABSTRACT

The increasing need for data driven decision making recently has resulted in the application of data mining in various fields including the educational sector which is referred to as educational data mining. The need for improving the performance of data mining models has also been identified as a gap for future researcher. In Nigeria, higher educational institutions collect various students' data, but these data are rarely used in any decision or policy making to improve the academic performance of students. This research work, attempts to improve the performance of data mining models for predicting students' academic performance using stacking classifiers ensemble and synthetic minority over-sampling techniques. The research was conducted by adopting and evaluating the performance of J48, IBK and SMO classifiers. The individual classifiers models, standard stacking classifier ensemble model and stacking classifiers ensemble model were trained and tested on 206 students' data set from the faculty of science federal university Dutse. Students' specific previous academic performance records at Unified Tertiary Matriculation Examination, Senior Secondary Certificate Examination and first year Cumulative Grade Point Average of students are used as data inputs in WEKA 3.9.1 data mining tool to predict students' graduation classes of degrees at undergraduate level. The result shows that application of synthetic minority over-sampling technique for class balancing improves all the various models performance with the proposed modified stacking classifiers ensemble model outperforming the various classifiers models in both performance accuracy and RSME values making it the best model.

Corresponding Author:

Ayuba John
Department of Cyber Security Federal,
University Dutse Jigawa State,
Dutse, Jigawa, Nigeria.
Email: ayuba.john@fud.edu.ng

1. INTRODUCTION

Decision making has gradually become data driven recently, due to the large amount of data available as a result of advancement in information and communication technology (ICT). Data mining has been applied in various fields like medical, marketing, machine learning, artificial intelligence, customer relations etc. Recently, Data mining is widely used on educational dataset which is referred to as educational data mining (EDM) and has now become a very useful research area [1]. This new emerging field, called educational data mining, [2] is concerned with developing methods that discovers knowledge in data originating from educational environments. To do this it uses different data mining techniques and machine learning algorithms. The study [3] indicates that some of the problems related to students' success in a course are hard to solve simply because usual statistical methods are not deep enough to discover the hidden patterns and knowledge, useful for educational processes planning and organization. Therefore, there is need to adopt data mining

technique for solving problems related to students' success using data originating from educational environments. Various data mining techniques have been implemented in research studies for educational data mining. The research work conducted [4] categorized all the various methods used in educational data mining into the following categories: Classification, Clustering, Relationship mining, Discovery with models and finally Distillation of data for human judgment. These data mining methods have been applied in many research works and were reported to have better performance than other methods. A cross validation test result [5] indicates that data mining techniques predicted significantly better than its statistical counterpart. Therefore, [6] suggested that with the increasing need for data mining and analyses, there is a need for improving the performance of data mining models and machine learning algorithms.

2. RELATED STUDIES

In recent time various research studies have been conducted on predicting students' academic performance using various data mining techniques and machine learning algorithms. The study of [7] adopted only two classifiers algorithms in predicting the dropout features of students. The result of the research on three different datasets which contains different students' attributes, such as: nationality of the students, sex, city of living, high school grades, program enrolled, number of earned credits in the first year of study and average grade in the first year of study indicates that data mining with J48 decision tree algorithm is more accurate than Naïve Bayes classifier algorithm with an accuracy of 81.1679 %. The research only considered two classifier algorithms. Meanwhile, [8] used more classifier algorithms which are; Neural Network (NN), Decision Tree, Support Vector Machine (SVM), K-nearest neighbor (KNN), Naïve Bayes and Rule Based to predict learners' progression in tertiary education. The finding from the research indicates that SVM has the highest performance accuracy of 73.33% and the least performance was recorded by Logistic Regression which has 60.05% accuracy. Only psychometric factors related to students are considered in conducting the research. Academic performance of student [2] is not a result of only one deciding factor besides it heavily hinges on various factors like personal, socio-economic, psychological and other environmental variables.

The work of [9] made use of three classifiers which are Naïve Bayes, Decision tree and Neural Network. In the study, continuous attributes were discretized using optimal equal width binning and Synthetic Minority Over-Sampling (SMOTE) technique was used to increase the volume of data, because there were limited instances in the acquired data during preprocessing. Neural Network and Naïve Bayes were reported to be more accurate than Decision tree for classification when Optimal Equal Width Binning and Synthetic Minority Over-Sampling (SMOTE) techniques are applied on data. Both classifiers have an accuracy of 71.6% though; Neural Network algorithm was discovered to be slower when compared to the Naïve Bayes algorithm thereby making Naïve Bayes model better than the neural network model. [2] 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, SMO, J48 and REPTree. The work shows that multilayer Perception has the highest accuracy of 75% and RepTree has the least accuracy with 67.76%.

A comparative analysis of three selected classification algorithms; Decision Tree (DT), Naïve Bayes (NB), and Rule Based (RB) was conducted by [10] to predict students' academic performance. The analysis was done to discover the best techniques to develop a predictive model for Student Academic Performance of first semester performance for first year Bachelor of computer science students at Universiti Sultan Zainal Abidin. Rule Based classifier was discovered to be the best model amongst the other classifiers by receiving the highest performance accuracy value of 71.3%. The model in this study does not provide detailed information about the students' performance. It only predicted the first-year performance of students not the graduation performance and classifies students' performance into poor, average and good. The research on early prediction of students' Grade Point Average (GPA) by [11] also showed that support vector machine (SVM) classifier algorithm prediction is more accurate than the extreme learning machine method and neural network. With performance accuracy of 93.06% when second year GPA of students is considered and 97.98% when the third year GPA of students are considered. Three supervised machine learning algorithms' performance were evaluated on students' assessment data characteristics by [14] to predict success in a course (either passed or failed) the result indicates that base on their prediction accuracy, ease of learning and user friendly characteristics, Naïve Bayes classifier outperforms decision tree and neural network classifiers.

The research conducted by [13] to predict students' performance shows that Random forest is a more accurate and faster algorithm compared to decision tree, K-Nearest Neighbour (IBK) and Multi-layer perceptron algorithms with an accuracy of 89.23%. Predicting academic performance of students is [14] challenging since students' academic performance depends on diverse factors such as personal, socio-economic, psychological and other environmental variables. The study also identify ensemble methods are the most influential development in Data Mining and Machine Learning in the past decade. An approach for predicting students' academic performance using ensemble model was presented in the study of [15]. Stacking

ensemble technique was used in [16] for predicting academic achievement of students. Performance of three classifiers algorithms were evaluated and stacking ensemble technique was used to combine the three classifiers and a better root mean square error (RMSE) value of 0.1291 was obtained as compare to 0.1898 for back propagation neural network, 0.1314 for M5P and 0.1343 for support vector machine.

Stacking is one of the ensemble techniques used by researcher with the aim of improving model's performance. [17] Stacking ensemble technique has capability of combining heterogeneous base classifiers and a Meta classifier is trained for final prediction. Predictions output of base classifiers are fed directly as data input into the Meta classifier for training and final prediction.

Therefore, in this research work, the improvement of the performance of stacking classifier ensemble model was considered, so that only instances that are correctly predicted from the base classifiers are fed to the meta classifier.

3. RESEARCH METHOD

The methodology of Educational Data Mining was 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 in this context [7]. But in general, it was stated in [18] that Data mining processes follows a set of steps that must be executed regardless of the algorithms or methodology that will be implemented. In this study the Cross Industry Standard Process for Data Mining (CRISP-DM) was adopted.

3.1. 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 student's data which represents 80% of the data set while 42 students' data which represent 20% of the data set was used in model testing.

3.2. 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, they have to be further pre-processed. The attribute filter in WEKA 3.9.1 was used to remove noisy and incomplete data. The final summary of attributes used for conducting the experiments after data cleaning are presented in Table 1.

Table 1. 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

3.3. Modeling

The model this research work attempts to develop is a stacking classifiers ensemble model. The data set for the study is small and imbalance as such, machine learning algorithms that are likely to perform well in developing this type of model based on previous studies were adopted. The dataset before class balancing as shown in Figure 1. 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 illustrated in Figure 2.

The various models were trained and tested using 10-fold cross validation to avoid over-fitting the models. Proposed model framework can be seen in Figure 3.

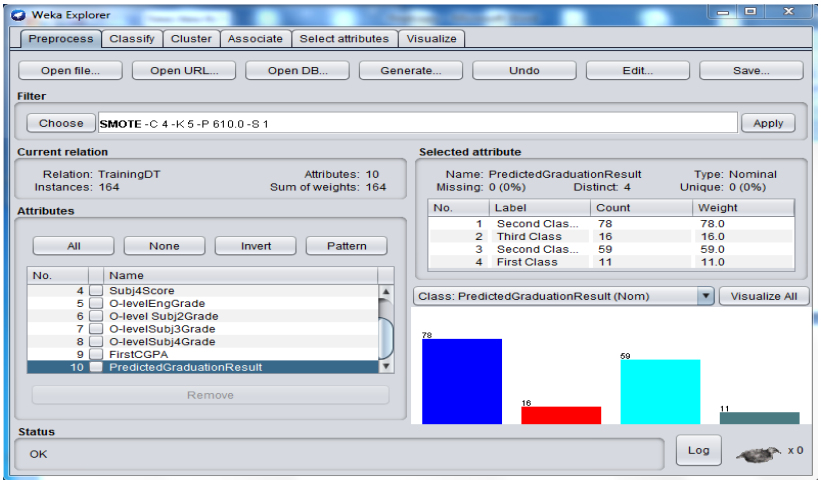

Figure 1. Dataset before class balancing

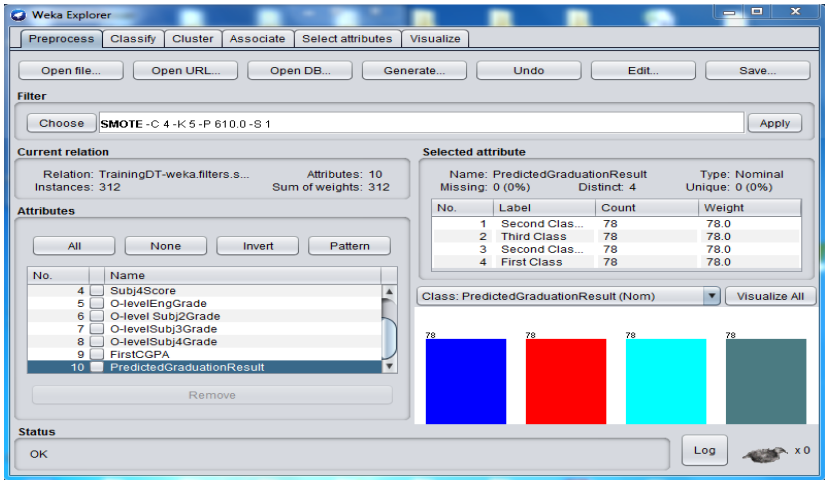

Figure 2. Dataset after class balancing with SMOTE on WEKA explorer

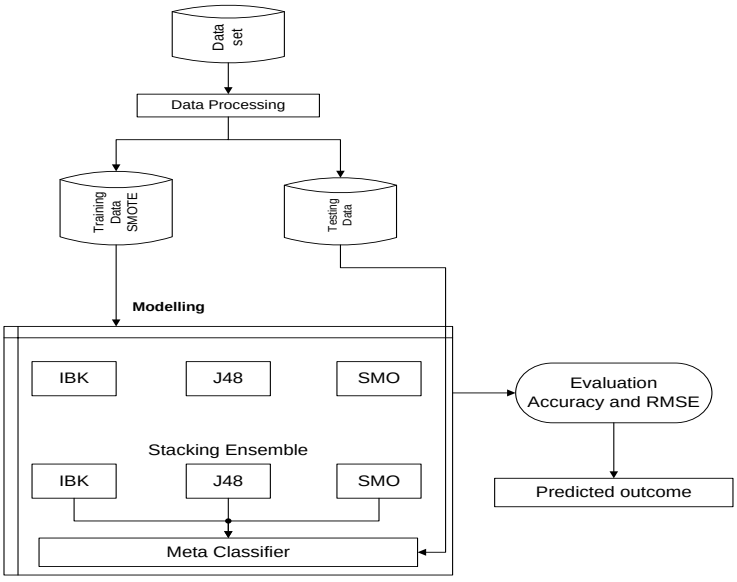

Figure 3. Proposed model framework

3.4. Model training and testing

In this research, series of training and testing were carried out on the using the various model by dividing 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. 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.

3.5. Model evaluation

To evaluate the performance of the various models, Performance accuracy and Root mean square error (RMSE) was used to indicate the various model performances which are presented in tabular forms.

4. RESULTS AND DISCUSSION

The results as obtained on training the various models using the training data set before class balancing is presented in Table 2 while the various models performance results after class balancing with SMOTE is presented in Table 3 The result from Table 3 indicates that class balancing using SMOTE results in improving all the various models performance. Though, all the various models recorded improvement in their performance. The proposed modified stacking classifiers ensemble model outperformed the other models in both performance accuracy and RMSE values which makes the model better than the other classifiers model.

Table 2. Performance of various model 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	Standard Stacking	87.1951%	0.2404	0.872	0.082	0.881
5	Modified Stacking	93.9024%	0.1719	0.939	0.048	0.941

Table 3. Performance result of various model 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	Standard Stacking	96.7949%	0.1098	0.968	0.011	0.969
5	Modified Stacking	97.7564%	0.1060	0.978	0.007	0.978

The various models performance accuracy results obtained on testing the various classifier models indicates that the modified stacking ensemble model outperformed the other models with an accuracy of 97.7564% and RMSE of 0.1060.

5. CONCLUSION

Data mining can be applied on students' data available to higher educational institutions to develop models for predicting students' graduation classes of degrees early using students' first year CGPA, UTME subjects' scores and their corresponding grades in SSCE. Resolving class imbalance problem in data set used for developing data mining models to predict students' academic performance using synthetic minority over-sampling technique (SMOTE) results in improving model performance.

REFERENCES

[1] B. Ryan & Y. Kalina, "The State of Educational Data Mining in 2009: A Review and Future Visions," *Journal of Educational Data Mining, Artiucle 1,* vol. 1, no.1, 2009.

[2] K. Parneet, M. Singh & G. S. Josan, "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,* pp. 500-508, 2015.

[3] N. Srecˇko & Z. Moti, "Student Data Mining Solution-Knowledge Management System Related to Higher Education Institutions," *Expert Systems with Applications, at ScienceDirect,* vol. 41 pp. 6400-6407. 2014.

[4] P. Nithya, B. Umamaheswari, A. Umadevi., "A Survey on Educational Data Mining in Field of Education," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, vol. 5, no. 1, pp.69-78 2016.
[5] S. A. Mohamed, W. Husain, N. A. Rashid., "A Review on Predicting Student's Performance using Data Mining Techniques," *SiecnceDirect Procedia Computer Science the Third Information Systems International Conference*, no. 72, pp. 414-422, 2015.
[6] M. S. Tabra & A. Lawan, "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)*, vol. 3, no 1, pp. 23-30, 2017.
[7] N. Vlatko, S. Riste etal., "Educational Data Mining: Case Study for Predicting Student Dropout in Higher Education," https://www.researchgate.net/publication/282333827 Conference Paper, April, 2015.
[8] G. Geraldine, C. McGuinness, P. Owende., "An Application of Classification Models to Predict Learner Progression in Tertiary Education," Conference Paper, February 2014 DOI: 10.1109/IAdCC.2014.6779384. 2014.
[9] J. S. Tanveer, R. I. Rashu etal., "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*, 2015.
[10] A. Fadhilah, N. H. Ismail, A. Abdul Aziz., "The Prediction of Students' Academic Performance Using Classification Data Mining Techniques," *Applied Mathematical Sciences*, vol. 9, no. 129, pp. 6415-6426, 2015. http://dx.doi.org/10.12988/ams.2015.53289
[11] Teressa T. & Chikohora, "A Study of the Factors Considered when Choosing an Appropriate Data Mining Algorithm," *International Journal of Soft Computing and Engineering (IJSCE)*, vol. 4, no. 3, pp.42-45, July 2014. ISSN: 2231-2307
[12] O. Edin & M. Suljić, "Data Mining Approach for Predicting Student Performance," *Economic Review – Journal of Economics and Business*, vol. x, no. 1, pp. 4-12, May 2012.
[13] M. S. Mythili & A. R. M. Shanavas, "An Analysis of students' Performance Using Classification Algorithms," *IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727*, vol. 6, no. 1, pp. 63-69, Ver. III Jan. 2014. www.iosrjournals.org. 2014
[14] S. Sajadin, M. Zarlis, D. Hartama, S. Ramliana, E. Wani, "Prediction of Student Academic Performance by an Application of Data Mining Techniques," *2011 International Conference on Management and Artificial Intelligence Ipedr*, vol. 6, pp. 110-114, 2011.
[15] R. Sikora & O. H. Al-laymoun, "A Modified Stacking Ensemble Machine Learning Algorithm Using Genetic Algorithms," *Journal of International Technology and Information Management*, vol. 23, no.1, pp. 1-12. 2014
[16] N. Chanamarn, K. Tamee, P. Sittidech., "Stacking Technique for Academic Achievement Prediction," 2016 International Workshop on Smart Info-Media Systems in Asia (SISA 2016), pp.14-17. 2016
[17] D. Saso & B. Zenko, "Is Combining Classifiers with Stacking Better than Selecting the Best One?," *Springer Journal of Machine Learning*, vol. 54, pp. 255-273, 2004.
[18] T. Ahmet., "Early Prediction of Students' Grade Point Averages at Graduation: A Data Mining Approach," *Eurasian Journal of Educational Research*, no. 54, pp. 207-226, 2014.

BIOGRAPHIES OF AUTHORS

Yusuf Abdulazeez received his BSc. degree in computer Science from Adamawa State University Mubi, Nigeria in 2011 and presently a post-graduate student at the department of Computer Science, Bayero University Kano, Nigeria. His research interest includes Data Mining, Artificial Intelligence and HCI.

Ayuba John received his B.Eng. Engineering Degree in Computer Engineering from University of Maiduguri, Nigeria in 2010, and M.Eng. Computer Engineering Degree from University of Benin, Nigeria, in 2017, He has worked as a transmission Engineer in National Control Centre (NCC); Transmission Company of Nigeria (TCN) in 2014 and he is a member of the Nigerian society of engineers (NSE), currently a lecturer from Federal University Dutse, Nigeria. His research interests are in the areas of Microelectronics' Intelligent Security System & Wireless Sensor Networks.