IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⪤ Impact Factor 7.105 ⪤ Vol. 9, Issue 7, July 2022

DOI: 10.17148/IARJSET.2022.9735

"Ensemble Machine Learning Models to Predict Students' Learning Behavior"

Susheelamma K H[1], Dr K M Ravikumar[2]

Research Scholar, S J C Institute of Technology, Chickballapur- 562101, Karnataka, India[1]

Director, The Oxford College of Engineering, Bengaluru – 560068, Karnataka, India[2]

Abstract: The number of e-learning platforms and blended learning environments is continuously increasing and has sparked a lot of research around improvements of educational processes. Here, the ability to accurately predict student performance plays a vital role [5]. Specialized scope classifiers are then combined to an ensemble to robustly predict student learning behavior on learning objectives independently of the students' individual learning setting [7]. The study aims to predict students learning behavior based on students' interactions with the virtual learning environments [6]. Firstly predictive models are built using traditional classifiers Logistic Regression, Support Vector Machine which are shown good performance. Later implemented popular ensemble methods based on bagging and boosting. The result revealed that the accuracy, precision, recall and F1 score had been considerably improved by ensemble bagging and boosting methods than that of the individual classifiers. The data used in the study is the Open University Learning Analytics dataset (OULAD) set [13] of year 2014.

Keywords: Classification, Machine learning, Virtual learning environment. Ensemble methods, Bagging, Boosting

I. INTRODUCTION

One of the popular fields of interest in the recent times is Educational Data Mining (EDM) [12], the data mining techniques have been used to extract useful knowledge from the available educational data [8]. The research also contributes towards extracting hidden useful information from the educational data. The previous works similar to the study had mainly focused on using single model for prediction [4] but in this research the ensemble model has been used for the prediction. The ensemble method combines the result of multiple individual models that can improve the reliability and performance of the model. Ensemble techniques have been very popular for predictive modeling almost in every field at recent time.

This paper mainly fulfill two objectives:

1) To compare and analyze the performance of the ensemble method in terms of various performance measures like Accuracy, Recall, Precision, and F-Measure.

2) To predict the learning behavior of open university students' with various attributes.

This paper data of OULAD students is used to develop a predictive model that can classify a student's learning behavior to predict students' assessment submission which classify into one of the two categories (Submits, Not Submits). This paper presents submission of the Assessment as learning behavior which is predicted through the OULAD VLE Clicks feature i.e., interaction with the VLE [9]. The OULAD dataset includes demographics, performance, VLE data. The possible attributes influencing in predicting students' learning behavior is VLE data [3]. The task was performed by using a supervised data mining technique i.e. Classification. The classification was first performed by using the two traditional classifiers Logistic Regression (LR), Support Vector Machine (SVM), and. After that, ensemble voting [2] was implemented by taking these two classifiers as base learners. Ensemble Methods provide classification accuracy by aggregating the prediction of multiple classifiers [1]. The ensemble method constructs a set of base classifiers from training data and performs classification by taking the vote on the predictions made by each classifier. In this model, for improving the classification accuracy, the bagging and boosting algorithm were used.

The contribution of work is as follows

- Presenting an efficient feature extraction model on real-time data for analyzing learning behavior of student for improving student performance.

IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁑ Impact Factor 7.105 ⁑ Vol. 9, Issue 7, July 2022

DOI: 10.17148/IARJSET.2022.9735

- Feature extracted are trained using two ensemble machine leaning models such as Random Forest (RF) and eXtreme Gradient Boosting (XGB) classification model [10], [11] for automatic prediction of students' learning behavior.
- Result are presented in terms of Accuracy, Precision, Recall, F-Measure.

The paper is organized as follows: In section II the proposed students' learning behavior model using ensemble machine learning algorithm is presented. Experimental studies are discussed in section III. Finally section IV the paper is concluded and future work of research is described.

II. ENSEMBLE MACHINE LEARNING MODELS TO PREDICT STUDENT LEARNING BEHAVIOR

This section presents method to predict students' learning behavior using ensemble machine learning models. The architecture of proposed students' learning behavior detection model using machine learning algorithm is shown in Figure 1. First, presents feature extraction model. Then, the extracted feature set are trained using ensemble machine learning algorithms namely RF and XGB.

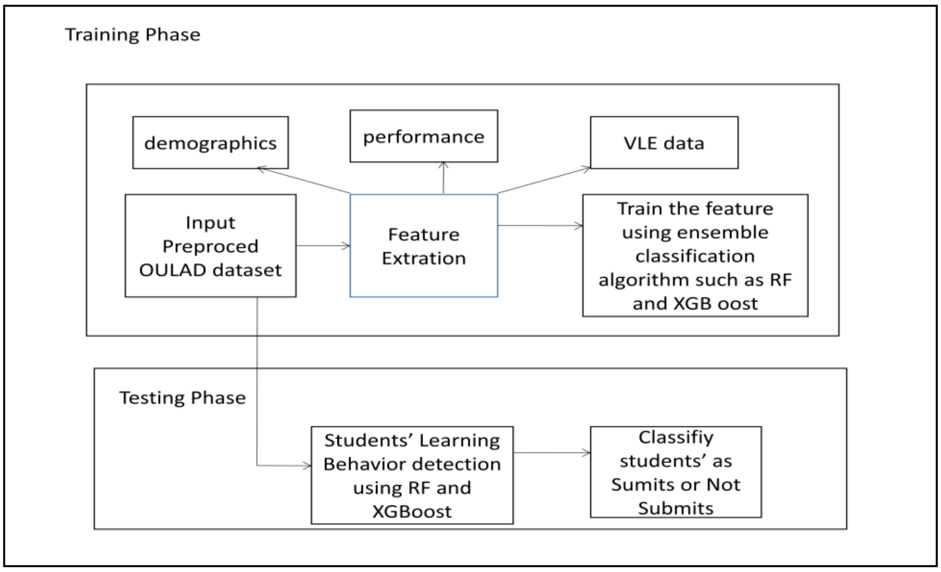

Figure 1 The proposed students learning behavior detection model using machine learning algorithm

A. Feature extraction

Further, it is noticed there is high correlation among feature sets because data is extracted utilizing same data source. Multicollinearity is confirmed by using Variance Inflation Factors (VIF is computed for correlation analysis for measuring linear correlation (dependency) among two parameters. The features having high VIF score are highly correlated. A threshold T is used for eliminating features (i.e, correlation value lesser than T are eliminated). Finally, obtained 4 feature including student ID also one of the feature. The entire data was then spiltted into training and testing data. The training data was imbalanced this could hamper the learning of the model. Therefore one of the resampling technique SMOTE was used to balance the majority and minority classes of training data. Synthetic Minority Over-sampling Technique (SMOTE) is an Over sampling technique that generates synthetic samples from the minority class [13]. It works by creating synthetic observations based upon existing minority observations.

The balanced training data was used for training each of the traditional classifiers i.e., LR, SVM. The performance measure of each model was evaluated using the unseen test data. In order to improve the performance ensemble boosting and bagging are used by taking these two classifiers as base learners. Bagging and boosting classifiers combine the predictions of base classifiers (LR and SVM) by averaging those predictions.

B. Classification using machine algorithm

The feature extracted in previous section are trained using four different machine learning methods such as LR, SVM, RF, XGB algorithms for detecting student learning behavior and predict Submission of Assessment. The data is divided

IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.105 ⁜ Vol. 9, Issue 7, July 2022

DOI: 10.17148/IARJSET.2022.9735

into train and test data and k-fold (10-fold) cross-validation was performed on them. Here the test data is composed of single occurrence of particular student and rest of the data are used during training phase. This is done for all the students and mean of result are obtained. The performance using LR, SVM, RF and XGB are compared. RF and XGB algorithms achieve much better performance than the LR and SVM which is experimentally proven below.

III. EXPERIMETAL RESULT AND ANALYSIS

This paper used the Scikit-learn Python library that has powerful tools to build the ML models and determined the accuracy of each algorithm using 10-fold cross-validation. This section present experiment analysis for detecting learning behavior of students' using machine learning technique such as LR, SVM, RF, XGB algorithm. The OULAD dataset is used for experiment analysis [12]. Figure 2 presents the number of clicks per activity, which indicates how much time the students spent on each activity.

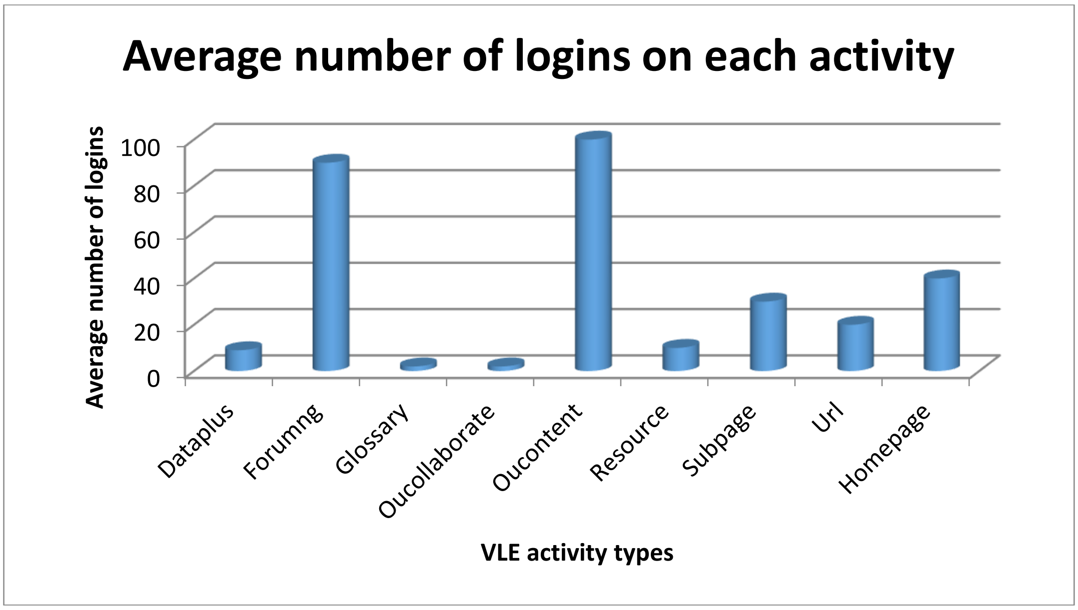

Figure 2 Number of clicks per activity

Figure 2 shows that the number of logins a student has for forumng and oucontent activities is greater than those for other activities. Additionally student forumng and oucontent engagement is greater when completing the first assessment. These findings demonstrate that forumng and oucontent have high importance in predicting students not submitting assessment.

Table 3 summarises the overall accuracy, Precision, Recall and F1-score for each of the algorithms on OULAD dataset. To identify low-engagement students (smaller number of clicks on activities) in the OU course, recall is paramount, but to identify high engagement students (those with a larger number of clicks on activities) then accuracy is important. Thus to identify students' of class not submitting assessment recall is important while identifying students' of class submitting assessment accuracy is important. LR, SVM, RF, XGB were selected as the appropriate classifiers for VLE data. LR achieved a recall for less active students' prediction of 0.81, a kappa value of 0.686, and an accuracy of 84.28%, and SVM achieved a recall of 0.8714 for low-engagement students, a kappa value of 0.757 and an accuracy of 87.84%, RF had a recall of 0.8778, a kappa of 0.807 and an accuracy of 90.34% in the current experiments.

Finally, XGB achieved a recall of 0.93 a kappa of 0.905, and an accuracy of 95.27%. These results indicate that the recall of XGBoost classification method is greater than that of the others models, which suggests that the performance of the XGBoost classifier in predicting low-engagement students is good compared to the alternatives. The accuracy of XGBoost classification method is greater than that of the others models, which suggests that the performance of the XGBoost classifier in predicting high-engagement students is also good compared to the alternatives.

Table 3 Accuracy, Precision, Recall and F-Measure of two class classification models build on LR, SVM, RF and XGB algorithms employed on OULAD dataset. The scores are sorted in descending order based on the F-Measure value and ranked them in ascending order.

Classifier name	Accuracy	Precision	Recall	F-Measure	Rank (F-Measure)
XGBoost	0.95	0.96	0.93	0.94	1
Random Forest	0.90	0.92	0.87	0.89	2
SVM	0.87	0.88	0.87	0.87	3
LR	0.84	0.88	0.81	0.84	4

Figure 14 shows the classification performance of LR, SVM, RF and XGB algorithms accuracy, Precision, Recall and F-Measure on OULAD dataset to identify students who are less active in VLE activities (smaller number of clicks on activities) in the OU course, to identify high active students (those with a larger number of clicks on activities).

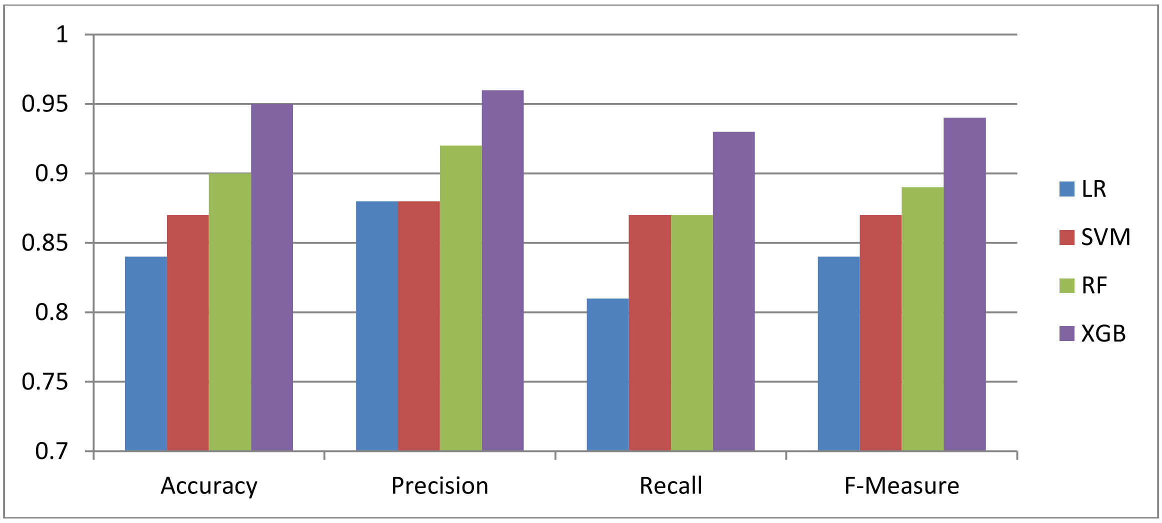

Figure 14 Comparison of performance using Accuracy, Precision, Recall, F-Measure for two class classification models build on LR, SVM, RF and XGB algorithms. The result shows the reliability of the proposed model. From the experimental results it is clear that boosting machine learning algorithms achieves good performance over bagging machine learning algorithms.

IV. EXPERIMETAL RESULT AND ANALYSIS

Designing classification methodologies using machine learning model for detecting students learning behavior is challenging problem. A cluster methodology doesn't yield good result when compared with supervised based methodologies. This paper presented efficient feature extraction method for detecting students' learning behavior level. Finally, the features extracted are trained using four machine learning model namely logistic regression, support vector machine and random forest and extreme gradient boosting algorithms. Finally, the model is tested considering k-fold cross-validation. From result attained it is seen accuracy score improvement RF and XGB over LR and SVM. From result achieved it is seen as the model obtain enough feature it is able to achieve higher accuracy. Future work would further consider building an improved classification model for students learning behavior.

REFERENCES

[1] Soobramoney, Ranjin, and Alveen Singh.,2019. "Identifying Students At- Risk With An Ensemble Of Machine Learning Algorithms." 2019 Conference on Information Communications Technology and Society (ICTA), 2019.
[2]Brownlee, J, "voting-ensembles-with-python" 2020, from machinelearningmastery: https:// machinelearningmastery.com/voting-ensembles-with-python/
[3] Kumari, Pooja, Praphula Kumar Jain, and Rajendra Pamula, "Review on Application of Data Mining Techniques for Student Performance Analysis", **Journal of Emerging Technologies and Innovative Research** (**JETIR**), 564-567, June 2019.

DOI: 10.17148/IARJSET.2022.9735

[4] F. Aman, A. Rauf, R. Ali, F. Iqbal and A. M. Khattak, "A Predictive Model for Predicting Students Academic Performance," 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA), 2019, pp. 1-4, doi: 10.1109/IISA.2019.8900760.

[5] S. Shishehchi, S. Y. Banihashem, N. A. M. Zin, and S. A. M. Noah, “Review of personalized recommendation techniques for learners in e-learning systems,” in Proceedings of 2011 International Conference on Semantic Technology and Information Retrieval, pp. 277–281, Putrajaya, Malaysia, June 2011.

[6] M. I. Olazabalaga, C. C. Garrido, and G. U. Ruiz, “Research on MOOCs: trends and methodologies,” Porta Linguarum, pp. 87–98, 2016.

[7] S. M. Jayaprakash, E. W. Moody, E. J. M. Lauria, J. R. Regan, and J. D. Baron. Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. Journal of Learning Analytics, pp. 6-47, 2014.

[8] Huang, S. & Fang, N., 2013. Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. Computers & Education, Issue 61, pp. 133-145.

[9] Hlosta, M. et al., 2014. Modelling student online behaviour in a virtual learning environment. Indianapolis, LAK 2014.

[10] T.Q. Chen, C. Guestrin, “XGBoost: a scalable tree boosting system”, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016.

[11] J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” Annals of statistics, pp. 1189-1232, 2001.

[12] Dr Brahmananda S.H. Dr. V. Cyril Raj, “Investigating the effect of An Integrated and flexible software system for teaching and learning ‘C’ programming language”, International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue 1, Jan-Feb 2012, pp.1065-1070.

[13] OULAD dataset available at: https://analyse.kmi.open.ac.uk/open_dataset

[14] Xiaojun Wu and Sufang Meng, "E-commerce customer churn prediction based on improved SMOTE and AdaBoost," 2016 13th International Conference on Service Systems and Service Management (ICSSSM), 2016, pp. 1-5, doi: 10.1109/ICSSSM.2016.7538581.