International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 5 Volume No.7 September - October – 2023
DOI:10.46647/ijetms.2023.v07i05.032 ISSN: 2581-4621

A Novel Approach to Predict Students Performance through Machine Learning

S Sukanya[1], Dr D William Albert[2]
[1]*M.Tech Student, Dept. of CSE, Ashoka Women's Engineering College, Kurnool, A.P*
[2]*Professor, Dept. of CSE, Ashoka Women's Engineering College, Kurnool, A.P*

Abstract— Forecasting student success can enable teachers to prevent students from dropping out before final examinations, identify those who need additional help and boost institution ranking and prestige. Machine learning techniques in educational data mining aim to develop a model for discovering meaningful hidden patterns and exploring useful information from educational settings. The key traditional characteristics of students (demographic, academic background and behavioral features) are the main essential factors that can represent the training dataset for supervised machine learning algorithms. The real goal is to have an overview of the systems of artificial intelligence that were used to predict Academic learning. This research also focuses on how to classify the most relevant attributes in student data by using prediction algorithm. Using educational machine learning methods, we could potentially improve the performance and progress of students more efficiently in an efficient manner. Students, educator and academic institutions could benefit and also have an impact.

In this paper, two predictive models have been designed namely students' assessments grades and final students' performance. The models can be used to detect the factors that influence students' learning achievement in MOOCs. The result shows that both models gain feasible and accurate results. The lowest RSME gain by RF acquire a value of 8.131 for students assessments grades model while GBM yields the highest accuracy in final students' performance, an average value of 0.086 was achieved.

Keywords— Massive Open Online Courses (MOOCs), Machine Learning, Receiver Operating Characteristic (ROC), Students Performance.

I. INTRODUCTION

There are many studies in the learning field that investigated the ways of applying machine learning techniques for various educational purposes. One of the focuses of these studies is to identify high-risk students, as well as to identify features which affect the performance of students.

With rapid advancements in technology, artificial intelligence has recently become an effective approach in the evaluation and testing of student performance in online courses.

Many researchers applied machine learning to predict student performance in [7], however few works have been done to examine the trajectories performance [8]. As a result, educators could not monitor the real-time student's learning curve.

Two sets of experiments are conducted in this work. In the first set of experiments, regression analysis is implemented for estimation of students' assessment scores. The student past and current activities in addition to past performance are employed to predict student outcome. In the second set of experiments, supervised machine learning method has been utilized to predict long-term student performance. Three types of candidate predictors have been considered firstly behavioral features, followed by temporal and demographic features. The proposed models offer new insight into determining the most critical learning activity and assist the educators in keeping tracking of timely student performance. To the best of our knowledge, student performance has been evaluated in online course using only two targets: "success" and "fail". Our model predicts the performance with three-class labels "success", "fail" and "withdrew".

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 5 Volume No.7 September - October – 2023
DOI:10.46647/ijetms.2023.v07i05.032 ISSN: 2581-4621

The rapid development of information technology (IT) has greatly increased the amount of data in different institutions. Huge warehouses contain a wealth of data and constitute a valuable information goldmine. This dramatic inflation in the amount of data in institutions has not kept pace with the efficient ways of investing these data. Thus, a new challenge has recently emerged, that is, transitioning from traditional databases that store and search for information only through questions asked by a researcher to techniques used in extracting knowledge by exploring prevailing patterns of data for decision making, planning and future vision.

II.SYSTEM ANALYSIS

Data Description

The OULAD dataset was captured from the Open University Learning Analytics Dataset (OULAD)repository. The open university in the UK delivers the online course in various topic for undergraduate and postgraduate students in the period between 2013-2014. The main composite table called “studentInfo” is linked to all tables.The"studentInfo "table includes information relevant to students’ demographic characteristics [15].

The information related to student’s performance is collected in “Assessments” and Student Assessment tables. The table “Assessments” contains information about the number, weight and the type of assessments required for each module. In general, each module involves a set of assessments, followed by the final exam. The assessments are Tutor Marked Assessment (TMA), Computer Marked Assessment (CMA).

The final average grade is computed with the sum of all assessments (50%) and final exams (50%). The “Student Assessment” table involves information relating to student assessment results, such as the date of the submitted assessment and the assessment mark[15].

The “Student Registration” table contains information about the date the students registered and unregistered in a particular module. The overall date is measured by counting numbers of unique days that students interact with courses until the course ends. In Open University online courses, students are able to access a module even before being a student of the course; however, it is not possible to access the course post-course closure date. The students' information related to their interaction with digital is store in learners Virtual Learning Environment dataset.

The field of machine learning has gained the attention of computer science and IT researchers. The data analysis field has become more essential than before, owing to the increasing amounts of huge data processed every day. The three basic types of machine learning are supervised, unsupervised and semi-supervised learnings [19]. In supervised learning, the training dataset only consists of labelled data. A supervised function is trained during the learning process, with the aim of predicting the future labels of unseen data. The two basic supervised problems are regression and classification, especially for discrete function classification and continuous regression [10]. Unsupervised learning aims to find meaningful, regular patterns without human intervention on unlabelled data. Its training set is made up of unlabelled data, and no instructor is present to help identify these patterns. Some popular supervised methods include clustering, novelty identification and dimensionality reduction [4, 11]. Semi-supervised learning is a combination of supervised and unsupervised learning processes. It is used to achieve enhanced results with few labelled examples. Its training dataset consists of labelled and unlabelled data. DT, NB, Logistic Regression, SVMs, KNN, SMO and Neural Network are well-known supervised techniques with accurate results in different scientific fields.

III.SYSTEM CONSTRUCTION

Simulation results- Final Students Performance Model

The classification analysis results for the second experiment presented as follows. The same set of machine learning classifiers in the previous experiment are used in this case study. As can be seen in a Table 3 all classifiers obtain similar ideal results, the highest performance achieved by Gbm with the value of 0.868 while RF, Nnet producing the value of 0.854, achieved the lowest accuracy.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 5 Volume No.7 September - October – 2023
DOI:10.46647/ijetms.2023.v07i05.032 ISSN: 2581-4621

Table 4 shows the class "Withdrawn" acquired the best accuracy of all Classifiers reaching an average value of 0.99 whereas the class "Fail" gives the lowest performance, with an approximate range of accuracy between 0.76-0.80.

The sensitivities are high overall classifiers for class "Withdraw" and "Pass". The best sensitivity achieved by Rpart reported the values of 0.99 and 0.92. The class "Fail" gained very low sensitivities across all classifiers. This is expected since the number of records with target class "Fail" are limited hence, the algorithm cannot learn well. With regards, to true negative instance, the Gbm and Nnet produce the best result, specificity =0.998 for class "Withdrawn". The poorest specificity gained by Rpart for class "Pass" obtained the values of 0.81. As can be seen, the best F1-Measure gained by Gbm yielded a value of 0.993, 0.864, 0.772, for the class "Withdrawn"," Pass" and "Fail" respectively. The lowest F1-Measure is shown for Rpart with the value of 0.67 over class "Fail".

ROC is used in this study to choose a decision threshold value for the true and false positive rate across each class. Overall, a range of AUC values between 0.99-0.82 for all classes was obtained. As previously mentioned, the demographic behavioral and temporal features in classification analysis were combined. In this model the total numbers of variables are 35. As a result, the predictive model may suffer from the over fitting issue. In this case, we compare classifiers results in terms of train and test error which could give an indication of the over fitting problem.

Figure 3 displays the result of over fitting evaluation. It can be observed that training and test error are low for all classifiers. The lowest test and train error was obtained by Gbm. The RF, Nnet obtained a similar test error with an approximate percentage of 14%. The training errors are slightly higher in these classifiers. The largest error was acquired by the Mlp model. Although all models fit well for most classifiers, Mlp suffers from overfitting.

Classifier	Accuracy
Mlp	0.858
RF	0.854
Rpart	0.862
Gbm	0.868
Nnet	0.854

Table 3 Accuracy Result for Final Students Performance model

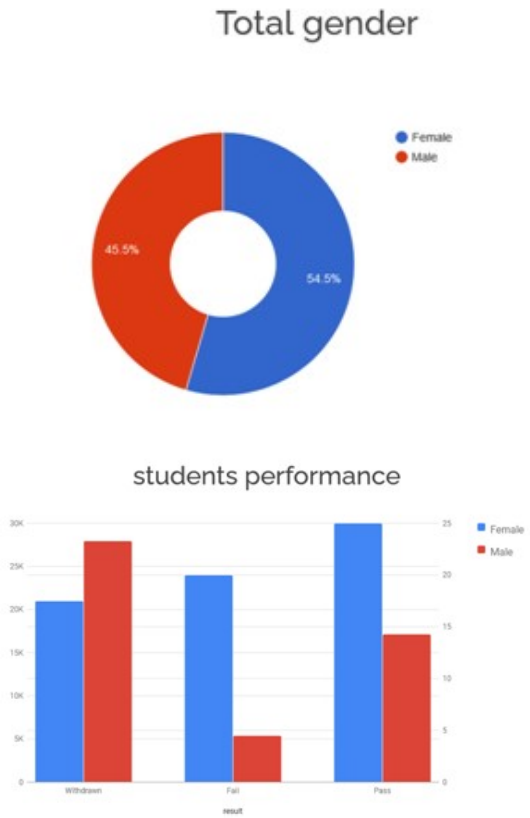
Table 4 Results for Final Students Performance Prediction Model

Classifier		**Performance Metrics**			
MLP	***ACC***	**F1**	**Sens**	**Spec**	**AUC**
Pass	0.858	0.850	0.892	0.824	0.916
Fail	0.782	0.690	0.631	0.932	0.886
Withdrawn	0.993	0.992	0.989	0.996	0.996
RF	***ACC***	**F1**	**Sens**	**Spec**	**AUC**
Pass	0.855	0.843	0.844	0.866	0.924
Fail	*0.808*	0.713	0.712	0.904	0.892
Withdrawn	0.995	0.993	0.991	0.990	0.995
Rpart	***ACC***	**F1**	**Sens**	**Spec**	**AUC**
Pass	0.866	0.865	0.923	0.810	0.867
Fail	0.767	0.671	0.582	0.952	0.821
Withdrawn	0.997	0.991	0.996	0.992	0.997
Gbm	***ACC***	**F1**	**Sens**	**Spec**	**AUC**
Pass	0.872	0.864	0.903	0.841	0.925
Fail	0.802	0.722	0.665	0.939	0.900
Withdrawn	0.994	0.993	0.991	0.998	0.997

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 5 Volume No.7 September - October – 2023
DOI:10.46647/ijetms.2023.v07i05.032 ISSN: 2581-4621

Nnet	*ACC*	F1	Sens	Spec	AUC
Pass	0.856	0.847	0.870	0.843	0.925
Fail	0.795	0.704	0.670	0.920	0.900
Withdrawn	0.994	0.993	0.991	0.998	0.998

ROC is used in this study to choose a decision threshold value for the true and false positive rate across each class. Figure 2 lists ROC curves. Overall, a range of AUC values between 0.99-0.82 for all classes was obtained. As previously mentioned, the demographic behavioral and temporal features in classification analysis were combined. In this model the total numbers of variables are 35. As a result, the predictive model may suffer from the overfitting issue. In this case, we compare classifiers results in terms of train and test error which could give an indication of the overfitting problem.



IV.CONCLUSION

Predicting student performance is important in the educational domain because student status analysis helps improve the performance of institutions. Different sources of information, such as traditional (demographic, academic background and behavioural features) and multimedia databases, are often accessible in educational institutions. These sources help administrators find information (e.g. admission requirements), predict the timetable scale of the class enrolment and help students decide how to choose courses depending on how well they will do in the chosen courses. The proposed model predicted student performance.

The final student performance predictive model revealed that student engagement with digital material has a significant impact on their success in the entire course. The findings' results also demonstrate that long-term students' performance achieves better accuracy than students' assessments grades prediction model, due to the exclusion of temporal features in regression analysis. The date of student deregistration from the course is a valuable predictor that is

DOI:10.46647/ijetms.2023.v07i05.032 ISSN: 2581-4621

significantly correlated with student performance. With the regression analysis, the data does not provide the last date of students' activity prior to undertaken assessments. The findings' results have been recommended to take into account the temporal features on predicting of subsequent assessments grades.

Future research direction involves the use of temporal features for predicting students' assessments grades model. With temporal feature time series analysis will be untaken, might be more advanced machine leering will be utilized.

V.REFERENCES

[1] J. Xu, K. H. Moon, and M. Van Der Schaar, "A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs," IEEE J. Sel. Top. Signal Process., vol. 11, no. 5, pp. 742–753, 2017.

[2] K. P. Shaleena and S. Paul, "Data mining techniques for predicting student performance," in ICETECH 2015 - 2015 IEEE International Conference on Engineering and Technology, 2015, no. March, pp. 0–2.

[3] A. M. Shahiri, W. Husain, and N. A. Rashid, "A Review on Predicting Student's Performance Using Data Mining Techniques," in Procedia Computer Science, 2015.

[4] Y. Meier, J. Xu, O. Atan, and M. Van Der Schaar, "Predicting grades," IEEE Trans. Signal Process. vol. 64, no. 4, pp. 959–972, 2016.

[5] P. Guleria, N. Thakur, and M. Sood, "Predicting student performance using decision tree classifiers and information gain," Proc. 2014 3rd Int. Conf. Parallel, Distrib. Grid Comput. PDGC 2014, pp. 126–129, 2015.

[6] P. M. Arsad, N. Buniyamin, and J. L. A. Manan, "A neural network students' performance prediction model (NNSPPM)," 2013 IEEE Int. Conf. Smart Instrumentation, Meas. Appl. ICSIMA 2013, no. July 2006, pp. 26–27, 2013.

[7] K. F. Li, D. Rusk, and F. Song, "Predicting student academic performance," Proc. - 2013 7th Int.Conf. Complex, Intelligent, Softw. Intensive Syst. CISIS 2013, pp. 27–33, 2013.

[8] G. Gray, C. McGuinness, and P. Owende, "An application of classification models to predict learner progression in tertiary education," in Souvenir of the 2014 IEEE International Advance Computing Conference, IACC 2014, 2014.

[9] N. Buniyamin, U. Bin Mat, and P. M. Arshad, "Educational data mining for prediction and classification of engineering students achievement," 2015 IEEE 7th Int. Conf. Eng. Educ. ICEED 2015, pp. 49–53, 2016.

[10] Z. Alharbi, J. . Cornford, L. . Dolder, and B. . De La Iglesia, "Using data mining techniques to predict students at risk of poor performance," Proc. 2016 SAI Comput. Conf. SAI 2016, pp. 523–531, 2016.

[11] B. Hore, s. Mehrotra, m. Canim, and m. Kantarcioglu, **"**secure multidimensional Range queries over outsourced data,**"** *vldb j.*, vol. 21, no. 3,Pp. 333**–**358, 2012.

[12] J. Mullan, "Learning Analytics in Higher Education," London, 2016.

[13] P. and K. Al-Shabandar, R., Hussain, A.J., Liatsis, "Detecting At-Risk Students With Early Interventions Using Machine Learning Techniques," *IEEE Access*, vol. 7, pp. 149464–149478, 2019.

[14] S. Jiang, A. E. Williams, K. Schenke, M. Warschauer, and D. O. Dowd, "Predicting MOOC Performance with Week 1 Behavior," in *Proceedings of the 7th International Conference on Educational Data Mining (EDM)*, 2014, pp. 273–275.

[15] L. Analytics and C. Exchange, "OU Analyse : Analysing at - risk students at The Open University," in *in Conference, 5th International Learning Analytics and Knowledge (LAK) (ed.)*, 2015, no. October 2014.

[16] R. Alshabandar, A. Hussain, R. Keight, A. Laws, and T. Baker, "The Application of Gaussian Mixture Models for the Identification of At-Risk Learners in Massive Open Online Courses," in *2018 IEEE Congress on Evolutionary Computation, CEC 2018 - Proceedings*, 2018.

DOI:10.46647/ijetms.2023.v07i05.032 ISSN: 2581-4621

[17] J.-L. Hung, M. C. Wang, S. Wang, M. Abdelrasoul, Y. Li, and W. He, "Identifying At-Risk Students for Early Interventions—A Time-Series Clustering Approach," *IEEE Trans. Emerg. Top. Comput.*, vol. 5, no. 1, pp. 45–55, 2017.
[18] C. Yun, D. Shin, H. Jo, J. Yang, and S. Kim, "An Experimental Study on Feature Subset Selection Methods," *7th IEEE Int. Conf. Comput. Inf. Technol. (CIT 2007)*, pp. 77–82, 2007.
[19] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16–28, 2014.