Predicting Student Success in Online Learning Environment Using Machine Learning

Puppal Gourav[1], Gogula Yagna Pavan[2] , Ramavath Ashok Kumar[3] ,Mrs. G. Kiranmai[4]

[1,2,3]UG Scholar , Dept. of AI & ML, St.Martin's Engineering College,Secunderabad , Telangana,India, 500100

[4]Assistant Professor, Dept. of AI & ML, St.Martin's Engineering College,Secunderabad, Telangana,India, 500100

puppalgourav@gmail.com

Abstract:

Recent years have witnessed an increased interest in online education, both massive open online courses (MOOCs) and small private online courses (SPOCs). This significant interest in online education has raised many challenges related to student engagement, performance, and retention assessments. With the increased demands and challenges in online education, several researchers have investigated ways to predict student outcomes, such as performance and dropout in online courses. This project presents a comprehensive review of state-of-the-art studies that examine online learners' data to predict their outcomes using machine and deep learning techniques. The contribution of this study is to identify and categorize the features of online courses used for learners' outcome prediction, determine the prediction outputs, determine the strategies and feature extraction methodologies used to predict the outcomes, describe the metrics used for evaluation, provide a taxonomy to analyze relatedstudies, and provide a summary of the challenges and limitations in the field.

Keywords:Massive open online course (MOOCs), Small PrivateOnline Course(SPOCs), IoT based industries, retention assessment.

1.INTRODUCTION

Online education has revolutionized the way people learn and has made education more accessible and affordable to numerous people worldwide. Despite the advantages and increased interest in online and distance learning, educational institutions are becoming increasingly concerned about students' performance and retention rates, particularly low certification/graduation and dropout/completion rates. Failing or dropping out of an online course or program is often considered a key parameter by institutional authorities for assessing program/course quality and allocating resources. Dropout and low certification rates can also pose a potential risk to an institution's reputation, profit, and funding. These outcomes also have significant consequences for a student's self-esteem, well-being, employment, and chances of graduating. As a result, finding more efficient approaches to forecasting students' performance as early as possible is critical forinstitutions, students, and educators to take proactive steps towardimproving students' online learning experiences and establishing intervention strategies that target students' needs. With the increased interest in online education and the large amount of data produced by learners through their interactions with online platforms, researchers have proposed methods to analyse learners' behavioural data to predict and improve educational outcomes. Learning analytics (LA), more commonly known as educational data mining (EDM), the task ofimproving students' online learning experiences and establishing intervention strategies that target students' needs. With the increased interest in online education and the large amount of data produced by learners through their interactions with online platforms, researchers have proposed methods to analyse learners' behavioural data to predict and improve educational outcomes. Learning analytics (LA), more commonly known as educational data mining (EDM), the task of analysing and finding patterns in learners' data for decision-making purposes, has attracted many researchers in recent years. Learning analytic tools enable institutions to gain an understanding of their students' status, actions and preferences individually, and in relation to their peers and the targeted educational objective. This allows the tailoring of material for individual students based on the projected outcomes and preferred learning styles. In online education, LA systems assess students' learning behaviour by utilizing extensive data collection of learners' data, including student enrolment information, past and current academic records, students' online behaviour, student surveys via questionnaires concerning courses and teaching techniques, and data from online discussion forums. Scholars have also examined various learning-behaviour attributes to predict learning outcomes, such as learners' performance and retention. To predict and analyse students' outcomes in online courses, researchers have examined several machine learning models, including support vector machines (SVMs), linear regression (LR), random forest (RF), and deep learning models such as convolutional neural networks (CNNs) and long short-term memory (LSTM).Recent years have witnessed an increased interest in online education, both massive open online courses (MOOCs) and small private online courses (SPOCs). This significant interest in online education has raised many challenges related to student engagement, performance, and retention assessments. With the increased demands and challenges in online education, several researchers have investigated ways to predict student outcomes, such as performance and dropout in online courses. This paper presents a comprehensive review of state-of-the-art studies that examine online learners' data to predict their outcomes using machine and deep learning techniques. The contribution of this study is to identify and categorize the features of online courses used for learners' outcome prediction, determine the prediction outputs, determine the strategies and feature extraction methodologies used to predict the outcomes, describe the metrics used for evaluation, provide a taxonomy to analyse related studies, and provide a summary of the challenges and limitations in the field.

2. LITERATURE SURVEY

[1] "Some studies in machine learning using the game of checkers"

Two machine-learning procedures have been investigated in some detail usi!Jg the game of checkers. Enough work has been done to verify the fact that a computer can be programmed so that it will learn to play better game of checkers than can be played by the person who wrote the program. Furthermore, it can learn to do this in a remarkably short period of time (8 or 10 hours of machine-playing time) when given only the rules of the game, a sense of direction, and a redundant and incomplete list of parameters which are thought to have something to do with the game, but whose correct signs and relative weights are unknown and unspecified. The principles of machine learning verified by these 'experiments are, of course, applicable to many other situations.

[2] "Machine learning"

Today, all institutions and companies are accelerating the use of AI technologies in their businesses to achieve a clear vision and quality results. The education sector is one of the sectors where AI can be used because of big data. In this work we created a machine-based learning model to predict a student's educational performance. The developed model relied on the student's previous data and performance in the last stage of the school. The model showed a very accurate accuracy rate that can be adopted.

[3] "Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades."

The computer programming course has always been considered a difficult course to get started, especially the teaching quality of computer programming courses in higher vocational schools is relatively poor, which has been a problem and challenge for schools and teachers. This paper designs a teaching evaluation method of computer programming course in higher vocational colleges based on big data. Firstly, the paper analyses the application of big data in the computer programming course of higher vocational college, and confirms the promoting effect of big data on the teaching of computer programming course of higher vocational college. Based on the evaluation results, the author puts forward some strategies to improve the teaching quality. The research results provide valuable reference for the teaching of computer programming courses and the application of big data in higher vocational schools.

[4] "Preventing student dropout in distance learning using machine learning techniques."

Student dropout occurs quite often in universities providing distance education. The scope of this research is to study whether the usage of machine learning techniques can be useful in dealing with this problem. Subsequently, an attempt was made to identifying the most appropriate learning algorithm for the prediction of students' dropout. A number of experiments have taken place with data provided by the 'informatics' course of the Hellenic Open University and a quite interesting conclusion is that the Naive Bayes algorithm can be successfully used. A prototype web based support tool, which can automatically recognize students with high probability of dropout, has been constructed by implementing this algorithm.

[5] "An empirical study of three machine learning methods for spam filtering"

The increasing volumes of unsolicited bulk e-mail (also known as spam) are bringing more annoyance for most Internet users. Using a classifier based on a specific machine-learning technique to automatically filter out spam e-mail has drawn many researchers' attention. This paper is a comparative study the performance of three commonly used machine learning methods in spam filtering. On the other hand, we try to integrate two spam filtering methods to obtain better performance. A set of systematic experiments has been conducted with these methods which are applied to different parts of an e-mail. Experiments show that using the header only can achieve satisfactory performance, and the idea of integrating disparate methods is a promising way to fight spam.

[6]"Mining Moodle data to detect the inactive and low-performance students during the Moodle course"

In web-based learning systems such as massive open online course (MOOC) and modular object-oriented developmental learning environment (Moodle), monitoring the student's activities as well as predict the low-performance students is an important task because it enables the instructors to award the students when their activities level drops from normal activities levels as well as having lower grades. We used several machine learning (ML) classification and clustering techniques to extract the pattern from student data during completing the Moodle course; which enables the instructor to detect the low-performance student in advance before the examination. The experimental result shows that the fuzzy unordered rule induction algorithm (FURIA) classification technique achieves high accuracy in detecting inactive students as well as predicts the different categories of the student during the Moodle course. The K-means clustering is also able to group the inactive and active users and poorly performed users. The result demonstrates that our proposed system will be easily integrated to Moodle system to send alert to inactive and low- performance students while completing the course and build efficient education environment for the students.

[7] Introduction to Machine Learning

The goal of machine learning is to program computers to use example data or past experience to solve a given problem. Many successful applications of machine learning exist already, including systems that analyze past sales data to predict customer behavior, optimize robot behavior so that a task can be completed using minimum resources, and extract knowledge from bioinformatics data. The second edition of Introduction to Machine Learning is a comprehensive textbook on the subject, covering a broad array of topics not usually included in introductory machine learning texts. In order to present a unified treatment of machine learning problems and solutions, it discusses many methods from different fields, including statistics, pattern recognition, neural networks, artificial intelligence, signal processing, control, and data mining. All learning algorithms are explained so that the student can easily move from the equations in the book to a computer program. The text covers such topics as supervised learning, Bayesian decision theory, parametric methods, multivariate methods, multilayer perceptrons, local models, hidden Markov models, assessing and comparing classification algorithms, and reinforcement learning. New to the second edition are chapters on kernel machines, graphical models, and Bayesian estimation; expanded coverage of statistical tests in a chapter on design and analysis of machine learning experiments; case studies available on the Web (with downloadable results for instructors); and many additional exercises. All chapters have been revised and updated. Introduction to Machine Learning can be used by advanced undergraduates and graduate students who have completed courses in computer programming, probability, calculus, and linear algebra. It will also be of interest to engineers in the field who are concerned with the application of machine learning methods.

[8]"SPOC-MFLP: A multi-feature learning prediction model for SPOC students using machine learning"

Learning analysis is one of the most important applications of machine learning. Many studies have proposed solutions to learning performance prediction using online learning data. Unlike the previous studies, this paper analyses online learning environment and formalizes the problem of online learning prediction. Based on the formalization, a multi-feature based learning prediction model for SPOC is proposed, called SPOC-MFLP, which generalizes the prediction problem of SPOC learning including objective, constraints, system and algorithms. The proposed SPOC-MFLP could be extended for MOOC and other online learning forms. Principle components analysis is adopted to discover the correlations of students 'online multi features, and linear regression and deep neural network are used to predict the learning performance. The predicted results include specific scores or segmented grades of the final exam

of SPOC, as well as students 'future specialized courses. Experimental data are collected from a SPOC in Huzhou University for two years and the experimental results reveal that the proposed SPOC-MFLP performs well.

[9] "Predicting student achievement based on temporal learning behaviour in MOOCs"

With the development of data mining technology, educational data mining (EDM) has gained increasing amounts of attention. Research on massive open online courses (MOOCs) is an important area of EDM. Previous studies found that assignment-related behaviours in MOOCs (such as the completed number of assignments) can affect student achievement. However, these methods cannot fully reflect students' learning processes and affect the accuracy of prediction. In the present paper, we consider the temporal learning behaviours of students to propose a student achievement prediction method for MOOCs. First, a multi-layer long short-term memory (LSTM) neural network is employed to reflect students' learning processes. Second, a discriminative sequential pattern (DSP) mining-based pattern adapter is proposed to obtain the behaviour patterns of students and enhance the significance of critical information. Third, a framework is constructed with an attention mechanism that includes data pre-processing, pattern adaptation, and the LSTM neural network to predict student achievement. In the experiments, we collect data from a C programming course from the year 2012 and extract assignment-related features. The experimental results reveal that this method achieves an accuracy rate of 91% and a recall of 94%.

[10] "A semi-supervised regression algorithm for grade prediction of students in distance learning courses"

Applying data mining methods in the educational field has gained a lot of attention among researchers in recent years. Educational Data Mining has turned into an effective tool for uncovering hidden relationships in educational data and predicting students' learning outcomes. Several supervised methods have been successfully applied with the purpose of identifying students at risk of failing or of predicting their academic performance. Recently, the implementation of Semi-Supervised Learning (SSL) methods in the educational process indicated their superiority over the supervised ones. SSL is an emerging subfield of machine learning seeking to effectively exploit a small pool of labeled examples together with a large pool of unlabelled ones. On this basis, a small number of students' data from previous years may be used as the training set of a learning model to predict future outcomes of current students. A number of rewarding studies deal with the implementation of classification methods in the educational field in contrast to regression, which is deemed to be a slightly touched task. In this paper, a novel semi-supervised regression (SSR) algorithm is presented for predicting the final grade of undergraduate students in a distance online course. To the best of our knowledge there is no study dealing with the implementation of SSR methods in the educational field. A plethora of attributes related to students' characteristics, academic performance and interaction within the course online platform form the training set, while several experiments were carried out confirming the superiority of the proposed algorithm over familiar regression methods. The experiment results show that the predictive performance of the proposed algorithm is increasing significantly over time, achieving a MAE value of less than 1.2358 before the middle of the academic year, which provides the advantage of early warnings and interventions.

3. PROPOSED METHODOLOGY

In this proposed system, we are employing OULAD dataset to predict grade and score using various machine learning algorithms like Random Forest and gradient boosting. Each algorithm performance is evaluated using accuracy, precision, recall and FSCORE.

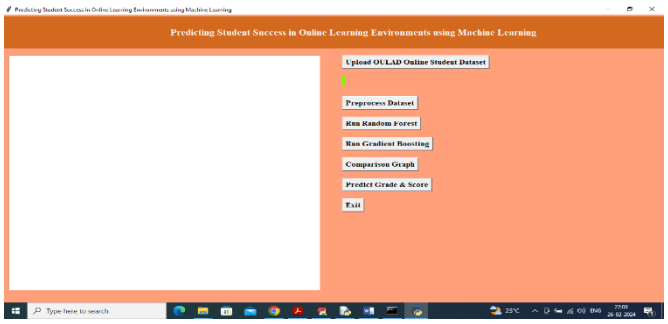
Random Forest classifier is used to predict grade and Random Forest Regressor is used to predict score. Similarly Gradient Boosting classifier to predict Grade and Gradient Boosting Regressor to predict score. Regressor algorithm performance is evaluated using RMSE (root mean square error). RMSE refers to difference between true value and predicted value so the lower the difference the better is the value.

Advantages:

1. High Accuracy
2. Takes less time
3. Early Identifications of struggling students.
4. Improved retention rates.

Applications:

1. **Higher Education** – Universities can use it to track student progress in online courses.
2. **E-learning Platforms** – EdTech companies like Coursera, Udemy, and Khan Academy can integrate it for personalized recommendations**.**
3. **Corporate Training** – Helps businesses monitor employee learning progress.



4. EXPERIMENTAL ANALYSIS

Fig. 1.Application Interface

In above screen click on 'Upload OULAD Online Student Dataset' button to upload dataset and get below page

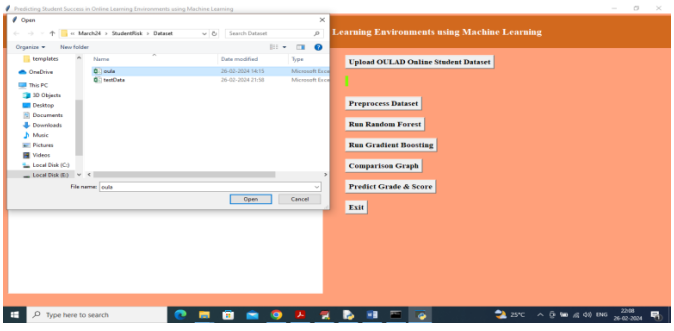


Fig.2.Uploading OULAD dataset

In above screen selecting and uploading OULAD dataset and then click on 'Open' button to load dataset and get below page

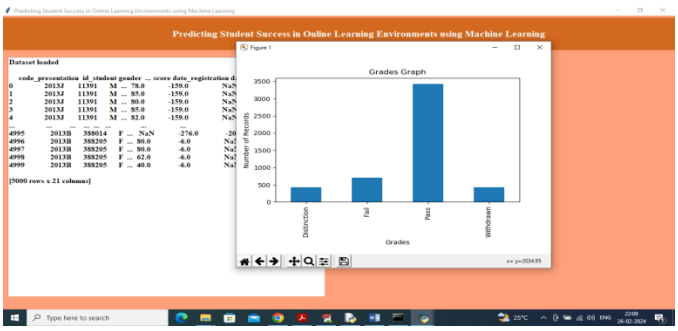


Fig.3.Grades Graph

In above screen dataset loaded and can see dataset contains both non-numeric and numeric values and by using processing technique will convert non-numeric to numeric values and in graph x-axis represents Grade and y-axis represents number of records found in that grade and now close above graph and then click on 'Pre-process Dataset' button to get below output

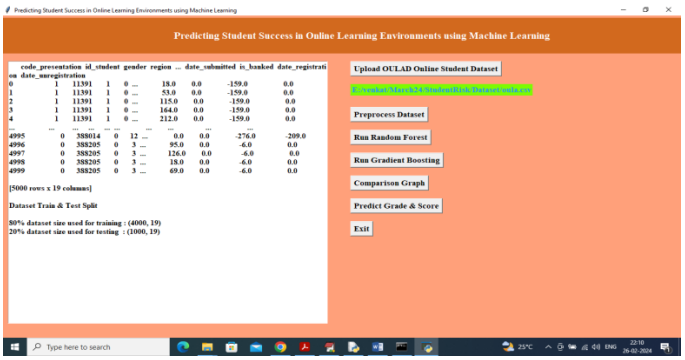


Fig.4.Running Random Forest

In above screen can see all values are converted to numeric format and then can see train and test data size and now click on 'Run Random Forest' algorithm button to get below output

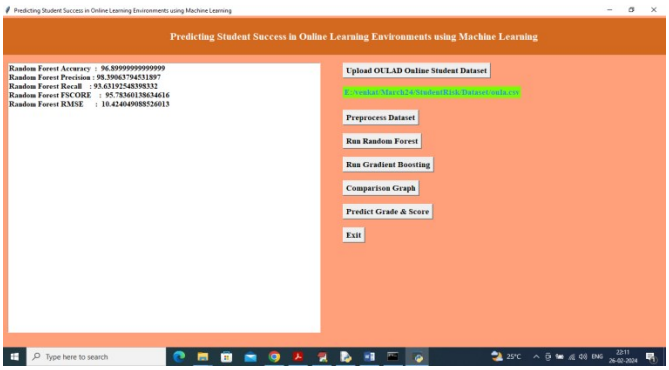


Fig.5.Random Forest Accuracy

In above screen can see Random Forest accuracy as 96% and RMSE as 10% and can see other metrics like precision, recall and FCSORE. Now click on 'Run Gradient Boosting' button to train boosting and get below output

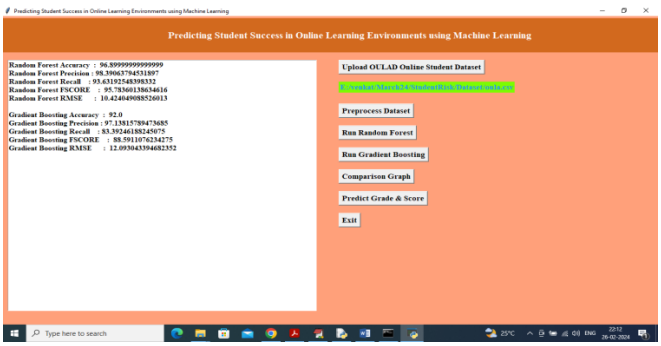


Fig.6. Random Forest and Gradient Boosting

In above screen gradient boosting accuracy is 92% and RMSE is 12% and now click on 'Comparison Graph' button to get below graph

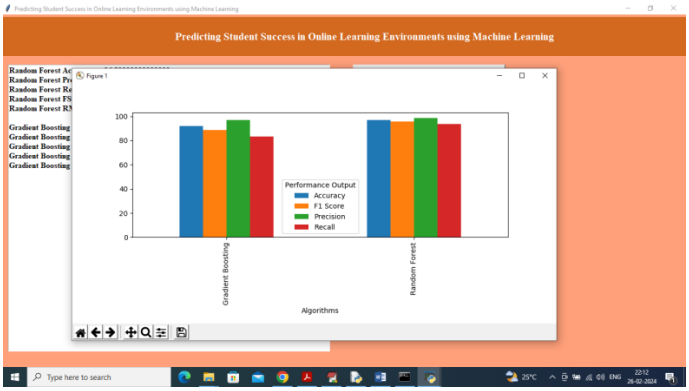


Fig.7.Comparison Graph

In above graph x-axis represents algorithm names and y-axis represents accuracy and other metrics in different colour bars and in both algorithm Random Forest got high accuracy and less RMSE. Now click on 'Predict Grade & Score' button to upload test data and get below output

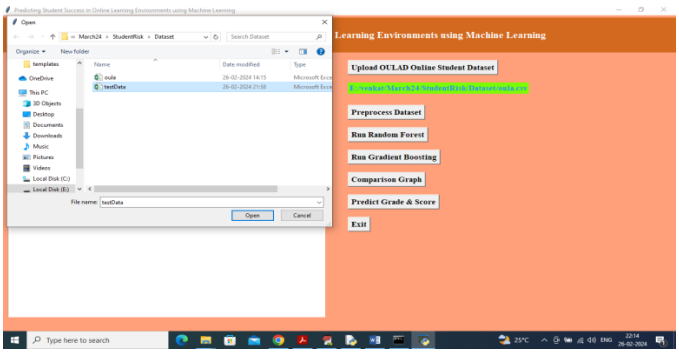


Fig.8.Uploading Test Data

In above screen selecting and uploading test data and then click on 'Open' button to get below prediction output

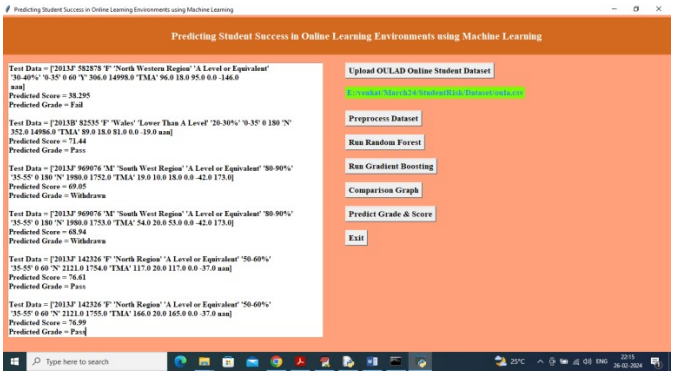


Fig.9.Predicting Grade and Score

In above screen in square bracket we can see Test data values and then in next two lines can see Predicted score and predicted Grade.

5. CONCLUSION

With the emergence of MOOCs and the expansion of online education in recent years, the prediction of learners' outcomes in online environments has attracted considerable attention. The spread of COVID-19 has also aided the growth of SPOCs, blended education, and an interest in monitoring student engagement and performance. Therefore, this study reviewed current strategies for predicting online-student outcomes in MOOCs and SPOCs. It summarized the predictive variables, online learning platforms, feature extraction, selection techniques, evaluation metrics, and the predictive models employed in this area. It also provided a thorough analysis and taxonomy for related research. Throughout our analysis, we found that most studies in the field utilized statistical features such as the number of downloaded materials and duration of video watching in a given time period. A small number of studies examined statistical temporal and raw temporal features in predicting learner outcomes. Studies conducted on benchmark datasets showed that statistical temporal features provide better results than raw features. Thus, further investigation of temporal features will provide a valuable understanding of users' learning progress and, eventually, their learner outcomes. Most temporally based LSTM or GRU models learners' time-series features. Further investigation of other recent sequence-based models, such as the attention-based model, is required. Studies using one-hot encoding to represent raw features and different representation techniques for raw features are worth investigating. Different machine-learning and deep-learning models have been used to predict learners' outcomes. RF and ANN are among the most effective machine learning models' performance and dropout prediction, whereas the sequence-based model provides the best performance on the publicly available dropout dataset. Further investigation of deep-learning models is recommended to predict student performance.

REFERENCES

[1]Arce, M.E.; Crespo, B.; Miguel-Álvarez, C. Higher Education Drop-out in Spain–Particular Case of Universities in Galicia. Int. Educ. Stud. 2015, 8, 247–264.

[2] Xavier,M.;Meneses, J. Dropout in Online Higher Education: A Scoping Review from 2014 to 2018; eLearn Centre, Universität Oberta de Catalunya: Barcelona, Spain, 2020.

[3] Baker, R.S.; Invent ado, P.S. Educational Data Mining and Learning Analytics. In Learning Analytics: From Research to Practice; Larusson, J.A., White, B., Eds.; Springer: New York, NY, USA, 2014; pp. 61–75.

[4] Moreno-Marcos, P.M.; Alario-Hoyos, C.; Maoz-Merino, P.J.; Kloos, C.D. Prediction in MOOCs: A Review and Future Research Directions. IEEE Trans. Learn. Technol. 2019, 12, 384–401.

[5] Ranjeeth, S.; Patchouli, T.; Paul, P.V. A survey on predictive models of learning analytics. ProcediaComputer. Sci. 2020, 167, 37–46.

[6] Hamim, T.; Benab Bou, F.; Sael, N. Survey of Machine Learning Techniques for Student Profile Modelling. Int. J. Emerg. Technol. Learn. 2021, 16, 136–151.

[7] Prianka, B.; Velardi, P.; Stilo, G.; Distant, D.; Ferally, S. A survey of machine learning approaches for student dropout prediction in online courses. ACM Computer. Surd. (CSUR) 2020, 53, 1–34.

[8] Gardner, J.; Brooks, C. Student success prediction in MOOCs. User Model. User-Adapt. Interact. 2018, 28, 127–203.

[9] Katara, R.; Gaba, J.; Garg, A.; Verma, V. A review on machine learning based student's academic performance prediction systems. In Proceedings of the 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 25–27 March 2021; pp. 254–259.

[10] Filius, R.M.; Bijl, S.G. Teaching Methodologies for Scalable Online Education. In Handbook for Online Learning Contexts: Digital, Mobile and Open; Springer: Berlin/Heidelberg, Germany, 2021; pp. 55–65.

[11] Amirah, E.A.; Hamsini, T.M.; Ajara, I. Mining Educational Data to Predict Student's academic Performance using Ensemble Methods. Int. J. Database Theory Appl. 2016, 9, 119–136.

[12] Rahman, M.H.; Islam, M.R. Predict Student's Academic Performance and Evaluate the Impact of Different Attributes on the Performance Using Data Mining Techniques. In Proceedings of the 2017 2nd International Conference on Electrical Electronic Engineering (ICEEE), Rajdhani, Bangladesh, 27–29 December 2017; pp. 1–4.

[13] Kozelek, J.; Hosta, M.; Rahal, Z. Open university learning analytics dataset. Sci. Data 2017, 4, 170171. [

[14] Adnan, M.; Habib, A.; Ashraf, J.; Mus Sadiq, S.; Raza, A.A.; Abid, M.; Bashir, M.; Khan, S.U. Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models. IEEE Access 2021, 9, 7519–7539.

[15] Jha, N.; Georgescu, I.; Moldovan, A. OULAD MOOC Dropout and Result Prediction using Ensemble, Deep Learning and Regression Techniques. In Proceedings of the 11th International Conference on Computer Supported Education, Heraklion, Greece, 2–4 May 2019; Scoters: Setubal, Portugal, 2019; Volume 2, pp. 154–164.

[16] Song, X.; Li, J.; Sun, S.; Yin, H.; Dawson, P.; Doss, R.R.M. SEPN: A Sequential Engagement Based Academic Performance Prediction Model. IEEE Intel. Syst. 2021, 36, 46–53.

[17] Hussain, M.; Zhu, W.; Zhang, W.; Abidi, R. Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores. Computer. Intel. Neurosis. 2018, 2018, 6347186.

[18] Stanford, U. Centre for Advanced Research through Online Learning (CAROL). Available online,

[19] Mubarak, A.A.; Cao, H.; Hizam, I.M. Deep analytic model for student dropout prediction in massive open online courses. Computer. Electra. Eng. 2021, 93, 107271.

[20] Mouridi, Y.; Saigal, M.; El Katine, H.; Berrada Fathi, W. A machine learning-based methodology to predict learners' dropout, success or failure in MOOCs. Int. J. Web Inf. Syst. 2019, 15, 489–509.

[21] Mouridi, Y.; Saigal, M.; Berrada Fathi, W.; El Katine, H. A Machine Learning Based Approach to Enhance Mooch Users' Classification. Turk. Online J. Distance Educ. 2020, 21, 47–68.

[22] KDD. KDD Cup 2015. Available online.

[23] Lai, S.; Zhao, Y.; Yang, Y. Broad Learning System for Predicting Student Dropout in Massive Open Online Courses. In Proceedings of the 2020 8th International Conference on Information and Education Technology, Okayama, Japan, 28–30 March 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 12–17.

[24] Zhou, Y.; Xu, Z. Multi-Model Stacking Ensemble Learning for Dropout Prediction in MOOCs. J. Phys. Conf. Ser. 2020, 1607, 012004.

[25] Ho, A.D.; Reich, J.; Nesterov, S.; Seaton, D.T.; Mullaney, T.; Waldo, J.; Chuang, I. Harvard and MITx: The first year of Open Online Courses; Harvard and MITx Working Paper No. 1; Harvard University: Cambridge, MA, USA, 2014.

[26] Imran, A.; Dalipi, F.; Kastrati, Z. Predicting Student Dropout in a MOOC: An Evaluation of a Deep Neural Network Model. In Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence, Bali, Indonesia, 19–22 April 2019; pp. 190–195.

[27] Al-Sha bandar, R.; Hussain, A.; Laws, A.; Keight, R.; Lunn, J.; Radi, N. Machine learning approaches to predict learning outcomes in Massive open online courses. In Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 14–19 May 2017; pp. 713–720.

[28] Liu, K.F.-R.; Chen, J.-S. Prediction and assessment of student learning outcomes in calculus a decision support of integrating data mining and Bayesian belief networks. In Proceedings of the 2011 3rd International Conference on Computer Research and Development, Shanghai, China, 11–13 March 2011; Volume 1, pp. 299–303.

[29] Wu, W.H.; Jim Wu, Y.C.; Chen, C.Y.; Kao, H.Y.; Lin, C.H.; Huang, S.H. Review of Trends from Mobile Learning Studies: A Meta-Analysis. Computer. Educ. 2012, 59, 817–827.

[30] Chen, W.; Brinton, C.G.; Cao, D.; Mason-Singh, A.; Lu, C.; Chiang, M. Early Detection Prediction of Learning Outcomes in Online Short-Courses via Learning Behaviours. IEEE Trans. Learn. Technol. 2019, 12, 44–58.