International Conference on “Interdisciplinary Research in Technology & Management” (IRTM 2021)

Conference Proceedings

Edited by

Satyajit Chakrabarti
Director, IEM Kolkata, India

Rintu Nath
Scientist – F, Vigyan Prasar, Department of Science and Technology, Govt. of India

Pradipta Kumar Banerji
Dean (Management Studies), Institute of Engineering and Management, Kolkata

Sujit Datta
Institute of Engineering & Management, Kolkata, India

Sanghamitra Poddar
Institute of Engineering & Management, Kolkata, India

Malay Gangopadhyaya
Institute of Engineering & Management, Kolkata, India

First edition published 2022
by CRC Press
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

and by CRC Press
6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

CRC Press is an imprint of Informa UK Limited

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalogue entry has been requested.

ISBN: 978-1-003-20224-0 (ebk)

DOI: 10.1201/9781003202240

Chapter 91

Prediction of Student Success using Student Engagement with Learning Management System

Eluwumi Buraimoh
School of Computer Science and Applied Mathematics
The University of the Witwatersrand, Johannesburg, South Africa

Ritesh Ajoodha
School of Computer Science and Applied Mathematics
The University of the Witwatersrand, Johannesburg, South Africa

Kershree Padayachee
Science Teaching and Learning Centre
The University of the Witwatersrand, Johannesburg, South Africa

Abstract—There has been a surge in student failure rates in blended-learning courses in recent times, which has generated considerable research interests. Engagement is identified as one of the core metrics for measuring students' success or failure in any learning system. This study utilizes machine learning algorithms on students' log-file data collected from an LMS to predict student success and increase the students' throughput rates. The machine learning predictive models considered in this study are the Naïve Bayes Classifier, Decision Tree, Gradient Boosting Tree, Linear Logistic Regression, Random Forest, Multilayer Perceptron Neural Network, and Support Vector Machines. The results provide an automatic predictive model for early detection of students at risk of failing for timely instructor intervention. The result serves as a feedback tool on learning for an increase in student performance. The machine learning algorithms' performances were evaluated using accuracy, precision, recall, and ROC-AUC for the best performing predictive model.

Keywords—Student success, At-risk student, Machine Learning, Predictive models, Student engagement, Educational Data Mining

I. INTRODUCTION

Educational Data Mining (EDM) is an emerging research area, serving a variety of instructional goals within web-based educational systems. [1] provided instructional purposes of EDM as evaluation of learning and instructional design efficacy, designing adaptive environments for students based on their actual behavior, providing input to both students and teachers, and detecting abnormal learning habits in the system.

Academic success is a huge challenge for higher education institutions across the world. The authors in [2] and [3] identified data analytics for detecting students failing academically and increasing the throughput rate in universities across the globe. Student engagement is a core metric for measuring a student's success or failure in any learning system. Recently, the rate of failure in undergraduate blended-learning courses is increasing, especially in science courses [4]. The trend in blended-learning failure necessitated the investigation into the key engagement metrics that have a high correlation with a student's performance in a blended learning environment. Blended-learning is the combination of conventional classroom learning, and technology-aided instructions [5]. However, there are many challenges associated with blended-learning systems. [6] in their study identified a lack of motivation of students in different courses and their respective course materials as major drawbacks in technology aided learning.

Some of the earlier studies classified performances of students into outstanding, average, or below average [7]. The students in the bracket of below-average are eventually identified to fail the course. [6] also emphasized in their research that excellent learning can be achieved by tracking student engagement in an educational program through different practices, which ultimately helps minimize dropout rates. This study seeks to investigate the vital student engagement metrics that affect students' success in a course.

This study is motivated by the need to increase students' throughput rates in a blended-learning course [4]. The study also aims to reduce to the barest minimum the number of atrisk students in higher education through the predictive model.

This research would provide automated predictive models to the Learning Management System used by the institution to monitor at-risk students. The predictive models will also produce significant insight for the educational instructors to improve their teaching material and student performances.

Section II provides a comprehensive literature review of student success prediction, particularly in relation to student engagement in the learning management system. Section III outlines the research methodology, including data collection and pre-processing, machine learning models, and evaluation techniques. Section IV highlights the results and conclusion.

II. LITERATURE REVIEW

Investigating student success using student engagement in the Learning Management System (LMS) is vital in reducing the at-risk of failing students in higher institutions of learning and improving learning outcomes [8]. Researchers have made various attempts to study student success using engagement in blended-learning, and online learning environments [6]. The correlation between students' performance and engagement has also been extensively investigated using data mining, statistical analysis, and machine learning. Many studies revealed that students' performance has always been correlated with different LMS engagement measures and strongly associated with their success in the course [4], [6], [9].

[6] studied student success using engagement in online learning using four variables: initial assessment scores, the highest level of education, final examination score, and the total number of clicks on Virtual Learning Environment (VLE). The authors' findings also showed that learners' clicks on "forumng and oucontent" are significant in predicting student success. The activities on the forum discussion and course content access positively impact student engagement and examination final grade [6]. The authors in [6] tested six predictive models in their study: Decision Tree, J48 Decision Tree, Classification and Regression Tree (CART), JRIP Decision Rules, Gradient Boosting Trees (GBT), and Naïve Bayes Classifier (NBC) on three extracted types of data (demography, performance, and learning behavior). The results of their study showed that J48, GBT, DT, and JRIP performed better than NBC and CART with accuracy values of 88.52%, 86.45%, 85.91%, 83.27%, 82.93%, and 82.25% respectively.

[10] studied the relationship between student engagement and performance in the e-learning environment by considering nine engagement metrics that are both frequency-related and time-related using association rule from learners' event logs. It was revealed in their study that student features such as the number of logins, the number of content read, and the number of forum read influenced the quiz performance, which later resulted in a higher final grade in the course. [10] proposed that student engagement can be a predictor of academic performance due to the positive correlation engagement has on performance.

The authors in [11] predicted student performance with ensemble methods on three models: Artificial Neural Network, Decision Tree, and Naïve Bayes. The study shows a direct correlation between the learner's interaction with LMS and academic performance. All the three models used in [11] study achieved over 80% accuracy.

In summary, past researches using statistical analysis, data mining, and machine learning techniques by researchers yielded various propositions in the study of student success. This paper focuses more on predictive models to monitor the students on the institution's LMS for a swift instructor or administrator's hypothetical intervention, especially for at-risk students.

III. METHODOLOGY

This research provides automated predictive models to the Learning Management System used by the institution to monitor at-risk students based on their engagement. The predictive models also produce significant insight for the educational instructors to improve their teaching material and student performances.

A. Data Collection

In this study, we used students' academic performance dataset that is freely available on Kaggle to investigate student success. The data were obtained from the Kalboard 360 Learning Management System using the xAPI learner activity tracker tool. The xAPI monitors learning and learner activity, such as login time or page read. The dataset consists of 480 student records and 16 attributes. The attributes are divided into three: (1) Demographic features such as gender and nationality. (2) Academic background features such as educational stage, Grade Level, and section. (3) Behavioral features such as raised hands-on class, opening resources, answering surveys by parents, and school satisfaction [11].

B. Pre-processing

- The target variable is classified into three classes based on student grades: Low, Medium, and High. A score between 0 to 69 indicates Low, scores between 70 to 89 indicates Medium, and 90 to 100 score is High. The attributes were classified into three categories: (1) Demographic attributes, (2) Academic attributes, and (3) Behavioral attributes [11]. The classification of the attributes is represented in Table I.
- Re-sampling strategies to obtain a more balanced data distribution were applied to the dataset's imbalance problem. A random under-sampling technique of the data to equal counts was employed to avoid bias and the predictive models' poor performance.
- Information gain attribute evaluation evaluates the attribute values by calculating the entropy in relation to the rank. Information Gain gives the importance of the attributes.

C. Predictive Models

In this paper, seven predictive models were used in the prediction of student success. The models are Gradient Boosting Tree, Multilayer Perceptron, Support Vector Machines, Naïve Bayes, Logistic Regression, Decision Tree, and Random Forest.

- Gradient Boosting Tree: Gradient boosting is a machine learning technique that draws recognition to its speed and accuracy of prediction, particularly with massive and complicated data. It reduces the risk of over-fitting. It works by combining a learning algorithm to achieve an efficient learner from several weak learners that are concurrently related [6].
- Multilayer Perceptron: Multilayer Perceptron is an artificial neural network feed-forward class. It uses the supervised learning concept called back-propagation for training and includes a minimum of three processing layers: input, hidden, and output layers [12].
- Support Vector Machines (SVMs): SVMs are supervised learning models with associated learning algorithms that analyze the data used to interpret classification problems. It is a commonly used Educational Data Mining because it has high accuracy in prediction [13]. SVMs can effectively perform a non-linear classification using the kernel trick, mapping their inputs into high-dimensional function spaces. The architecture of SVMs used in this study is from [14].
- Naïve Bayes Classifier (NBC): For most predictive problems, this algorithm is the most pragmatic and most straightforward learning approach. The Naïve Bayes Classifier (NBC) is based on Bayes' theorem with strong independence assumptions between the features using a probabilistic approach [6]. NBC is efficient because it takes less processing time and less training data compared to most machine learning models. The classifier estimates the parameters of the probability distribution P(A/B) on the training set of features B (16 features represented in table II) given class A (Low, Medium, or High), and then compute the posterior probability of the testing set. This leads to classifying the testing set based on the largest computed probability. The probability can be represented mathematically as:

$$P(A/B) = \frac{P(B/A) * P(A)}{P(B)} \tag{1}$$

Logistic Regression: Logistic regression is one of the mostly used in Educational Data Mining analysis. It is used to model-dependent variables with the aid of independent variables. It is based on the calculation of the highest probability, and, according to this probability, the data observed should be the most probable.

- Decision Trees: A Decision Tree has a flowchart-like structure where each internal node tests an attribute. Each branch corresponds to the attribute value, and each leaf node assigns a classification (failed or passed). The tree is constructed from the dataset by deciding which attributes at the child nodes better divide input features. In this case, we are using the concept of information gain. If a node has minimal entropy (highest information gain), it is used as a split node. When a study seeks to determine which features are important in a student prediction model, a decision tree is important [6].
- Random Forest: The Random Forest is an associative learning algorithm. It consists of several randomly generated decision trees. These randomly generated decision trees are merged to achieve higher accuracy and reliable predictions. The performance of random forest is usually better than decision trees [13].

The seven predictive models will be assessed using a confusion matrix, and the accuracy of each model will be provided after a 10-fold cross-validation [13].

IV. RESULTS AND DISCUSSION

The results of the accuracies of the seven predictive models after the 10-fold cross-validation is shown in table III. Random Forest has the best performance compared to the remaining six models with an accuracy value of 73.19%. The high accuracy of random forest is due to the model's ensembling of the decision trees characteristic it possesses [13]. The Naïve Bayes model follows the random forest directly with 72.44% accuracy. The other models' performances for the student success prediction are 72.42%, 72.29%, 70.31%, 66.36%, and 64.79% for Gradient Boosting, Logistic Regression, Decision Tree, Support Vector Machines and Multilayer Perceptron, respectively. The least performing model is Multilayer Perceptron, and this poor performance is due to the long learning time, and poor interpretability [13].

Table I: The students' attributes classification.

Attribute Classification	Attribute	Explanation
Demographic Attributes	Gender Nationality Place of Birth Relation	Statistical data such as age, gender
Academic Attributes	Stage ID Grade ID Section ID Topic Semester	Data related to student academic activities
Behavioral Attributes	Raise Hands Visited Resources Announcement Views Discussion Parent Answering Survey Parent School Satisfaction Student Absent Days	Student Engagement with LMS

The Information Gain in table II shows the ranking of the set of attributes in descending order. The top five attributes/features are the most contributing attributes to the model predictions. These five attributes fall under behavioral attributes of students on LMS as shown in table I and [11] also specified the categorization of these attributes. The top five features are:

(i) Visited Resources: the number of times students visited the resources,

(ii) Student Absence Days: the number of absent days on the LMS,

(iii) Raised Hands: the number of times the student raised hands to ask questions or make contributions on LMS,

(iv) Announcement Views: the number of times students viewed the announcement placed on LMS and

(v) Parent Answering Survey: The parent response to the surveys offered by the school.

The other attributes on levels 6 to 16 are categorized as demographic and academic attributes respectively. This is an indication that students' engagement has a direct correlation on students' academic performances [6].

Figures 1 to 7 are the Receiver Operating Characteristic (ROC) curves. ROC is probability distribution of both the True Positive Rate (TPR) and True Negative Rate (TNR). The X-axis is a False Positive Rate (FPR) and the Y-axis is a True Positive Rate (TPR) [15]. The AUC scale is from 0 to 1. If the value is greater than 0.5, then the model is considered as a good model [15] The performances of the predictive models shows that AUC scores are above 0.5. Random Forest model has the highest AUC score of 0.9198. AUC curve is calculated by this formula:

$$AUC = \frac{1}{2}(TPR + TNR) \qquad (2)$$

Table II: A ranking of the information gain (entropy) for a set of features to predict the students' success.

Rank	Entropy	Attribute Name
1	0.45801	Visited Resources
2	0.39745	Student Absence Days
3	0.37337	Raised Hands
4	0.2578	Announcements View
5	0.1504	Parent Answering Survey
6	0.12773	Nationality
7	0.1261	Relation
8	0.12292	Place of Birth
9	0.11393	Discussion
10	0.10676	Parent School Satisfaction
11	0.07611	Topic
12	0.05178	Gender
13	0.04748	Grade ID
14	0.01182	Semester
15	0.01058	Stage ID
16	0.00703	Section ID

Table III: Predictive model accuracies after 10-fold cross-validation

Predictive Model	Accuracy
Random Forest	73.19%
Naïve Bayes	72.44%
Gradient Boosting	72.42%
Logistic Regression	72.29%
Decision Tree	70.31%
Support Vector	66.39%
Multilayer Perceptron	64.79%

TPR and TNR are used in AUC and ROC curve as follows:

$$TPR = \frac{TP}{TP + FN} \tag{3}$$

$$TNR = \frac{TN}{TN + FP} \tag{4}$$

$$FPR = \frac{FP}{FP + TN} \tag{5}$$

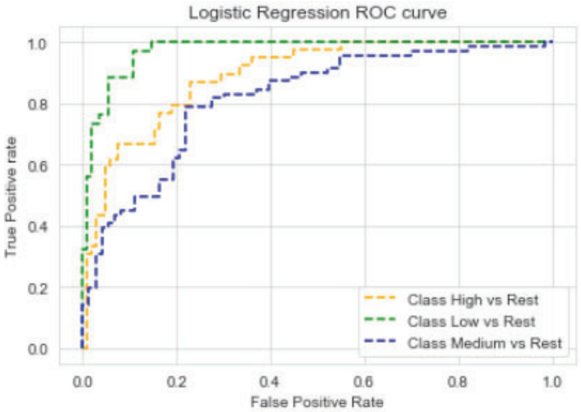

Figure 1: Logistic Regression Model ROC curve

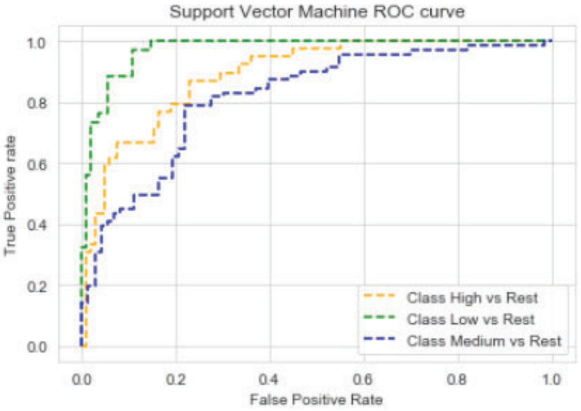

Figure 2: Support Vector Machines ROC curve

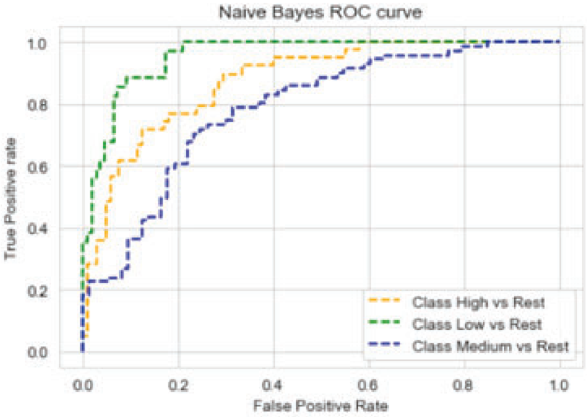

Figure 3: Naïve Bayes ROC curve

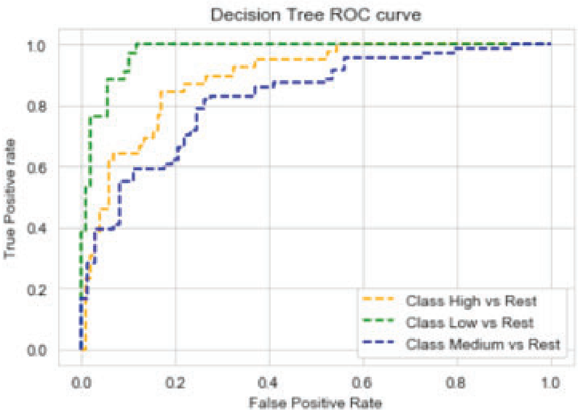

Figure 4: Decision Tree ROC curve

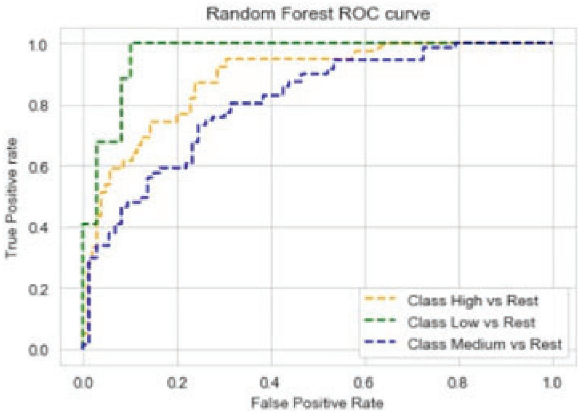

Figure 5: Random Forest ROC curve

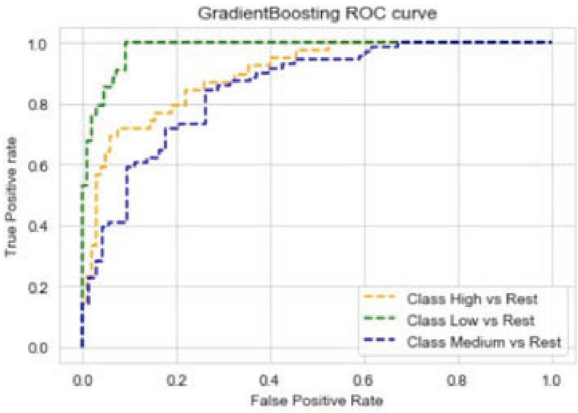

Figure 6: Gradient Boosting ROC curve

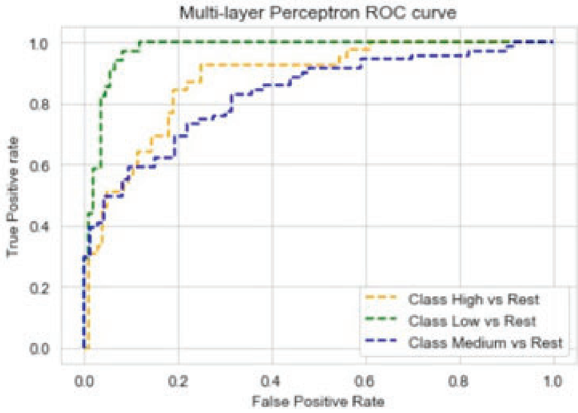

Figure 7: Multilayer Perceptron ROC curve

	Predicted			
		LOW	MED	HIGH
Actual	LOW	34	0	0
	MED	7	49	15
	HIGH	0	10	29

Figure 8: A Confusion Matrix describing the performance of the Logistic Regression predictive model.

	Predicted			
		LOW	MED	HIGH
Actual	LOW	32	2	0
	MED	15	38	18
	HIGH	32	14	23

Figure 9: A Confusion Matrix describing the performance of the Support Vector Machines predictive model.

	Predicted			
		LOW	MED	HIGH
Actual	LOW	31	3	21
	MED	11	39	21
	HIGH	0	7	32

Figure 10: A Confusion Matrix describing the performance of the Naïve Bayes predictive model.

	Predicted			
		LOW	MED	HIGH
Actual	LOW	34	0	0
	MED	0	64	7
	HIGH	0	2	37

Figure 11: A Confusion Matrix describing the performance of the Decision Tree predictive model.

		Predicted		
		LOW	MED	HIGH
Actual	LOW	34	0	0
	MED	4	57	10
	HIGH	1	0	38

Figure 12: A Confusion Matrix describing the performance of the Random Forest predictive model.

		Predicted		
		LOW	MED	HIGH
Actual	LOW	34	0	0
	MED	4	57	10
	HIGH	1	0	38

Figure 13: A confusion Matrix describing the performance of the Gradient Boosting Tree predictive model.

		Predicted		
		LOW	MED	HIGH
Actual	LOW	34	0	0
	MED	7	55	9
	HIGH	0	1	38

Figure 14: A confusion Matrix describing the performance of the Multilayer Perceptron predictive model.

V. CONCLUSION

Predictive models offer substantial insight in education to the instructor, administrator, and institution for student performance improvement, increase in throughput rate, and reduction in the dropout rate. The implication of this study is that the student will get the instructor or institution's timely intervention, thereby averting possible failure in their courses, increasing throughput rate, and other academic outcomes. The dataset used in this study is an LMS interaction of students on the Kalboard 360 Learning Management System. Seven models: Logistic Regression, Support Vector Machines, Naïve Bayes, Decision Trees, Random Forest, Gradient Boosting Tree, and Multilayer Perceptron Network were trained and evaluated. Random Forest performed better than the other models with an accuracy value and AUC score of 73.19% and 0.9198, respectively.

Behavioral attributes of the students' interactions with LMS were also identified to have a strong influence on model predictions. The ranking of the features shows that student engagement is positively correlated to student success or performance.

The limitation of this study is the size of the dataset used. A larger dataset would have given a better insight into the other influencing factors on student success prediction. The outcomes of this paper are based entirely on the study obtained in the data used. The contributions of this paper are: the provision of the vital engagement metrics on student's performances, early detection technique for instructors or lecturers to identify a student at risk of failing during courses for possible intervention based on the behavioral attributes and the impacts of student engagement on the students' success or performance. Based on the predictive models used in this paper and their evaluation metrics, we proposed that student engagement can play a vital role in predicting student success.

ACKNOWLEDGMENT

This work is based on the research supported in part by the National Research Foundation of South Africa (Grant numbers: 121835).

REFERENCES

[1] F. Castro, A. Vellido, A. Nebot, and F. Mugica, "Applying data mining techniques to e-learning problems," in *Evolution of teaching and learning paradigms in intelligent environment.* Springer, 2007, pp. 183–221.

[2] R. Ajoodha, A. Jadhav, and S. Dukhan, "Forecasting learner attrition for student success at a south african university," in *In Conference of the South African Institute of Computer Scientists and Information Technologists 2020 (SAICSIT '20), September 14-16, 2020, Cape Town, South Africa. ACM, New York, NY, USA, 10 pages.* ACM, 2020.

[3] A. D. Kumar, R. P. Selvam, and K. S. Kumar, "Review on prediction algorithms in educational data mining," *International Journal of Pure and Applied Mathematics,* vol. 118, no. 8, pp. 531–537, 2018.

[4] L. A. Buschetto Macarini, C. Cechinel, M. F. Batista Machado, V. Faria Culmant Ramos, and R. Munoz, "Predicting students success in blended learning—evaluating different interactions inside learning management systems," *Applied Sciences,* vol. 9, no. 24, p. 5523, 2019.

[5] W. W. Porter, C. R. Graham, K. A. Spring, and K. R. Welch, "Blended learning in higher education: Institutional adoption and implementation," *Computers & Education,* vol. 75, pp. 185–195, 2014.

[6] M. Hussain, W. Zhu, W. Zhang, and S. M. R. Abidi, "Student engagement predictions in an e-learning system and their impact on student course assessment scores," *Computational intelligence and neuroscience,* vol. 2018, 2018.

[7] M. Asiah, K. N. Zulkarnaen, D. Safaai, M. Y. N. N. Hafzan, M. M. Saberi, and S. S. Syuhaida, “A review on predictive modeling technique for student academic performance monitoring,” in *MATEC Web of Conferences,* vol. 255. EDP Sciences, 2019, p. 03004.

[8] F. Strydom, M. Mentz, and G. Kuh, “Enhancing success in south africa’s higher education: Measuring student engagement,” *Acta Academica,* vol. 42, no. 1, pp. 259–278, 2010.

[9] M. Hu and H. Li, “Student engagement in online learning: A review,” in *2017 International Symposium on Educational Technology (ISET).* IEEE, 2017, pp. 39–43.

[10] A. Moubayed, M. Injadat, A. Shami, and H. Lutfiyya, “Relationship between student engagement and performance in e-learning environment using association rules,” in *2018 IEEE World Engineering Education Conference (EDUNINE). IEEE,* 2018, pp. 1–6.

[11] E. A. Amrieh, T. Hamtini, and I. Aljarah, “Mining educational data to predict student’s academic performance using ensemble methods,” *International Journal of Database Theory and Application,* vol. 9, no. 8, pp. 119–136, 2016.

[12] C. A. C. Yahaya, C. Y. Yaakub, A. F. Z. Abidin, M. F. Ab Razak, N. F. Hasbullah, and M. F. Zolkipli, “The prediction of undergraduate student performance in chemistry course using multilayer perceptron,” in *IOP Conference Series: Materials Science and Engineering,* vol. 769, no. 1. IOP Publishing, 2020, p. 012027.

[13] T. Abed, R. Ajoodha, and A. Jadhav, “A prediction model to improve student placement at a south african higher education institution,” in *2020 International SAUPEC/RobMech/PRASA Conference. IEEE,* 2020, pp. 1–6.

[14] T. Park and C. Kim, “Predicting the variables that determine university (re-) entrance as a career development using support vector machines with recursive feature elimination: The case of south korea,” *Sustainability,* vol. 12, no. 18, p. 7365, 2020.

[15] A. J. Bowers and X. Zhou, “Receiver operating characteristic (roc) area under the curve (auc): A diagnostic measure for evaluating the accuracy of predictors of education outcomes,” *Journal of Education for Students Placed at Risk (JESPAR),* vol. 24, no. 1, pp. 20–46, 2019.