Application of Machine Learning Techniques to the Prediction of Student Success

Eluwumi Buraimoh
School of Computer Science
and Applied Mathematics
The University of the Witwatersrand
Johannesburg, South Africa
2287804@students.wits.ac.za

Ritesh Ajoodha
School of Computer Science
and Applied Mathematics
The University of the Witwatersrand
Johannesburg, South Africa
ritesh.ajoodha@wits.ac.za

Kershree Padayachee
Centre for Learning, Teaching and Development
The University of the Witwatersrand
Johannesburg, South Africa
kershree.padayachee@wits.ac.za

***Abstract*—This study presents six machine learning models in the prediction of student success in a technology-mediated environment. Student behavioral attributes with a learning management environment have proven to be a significant determinant in forecasting students' performance. This study attempts to provide the model with optimum accuracy to determine students who need assistance to improve their educational performances and other learning outcomes. We examined the impacts of SMOTE data re-sampling and the effect of attribute selection in this study. The models' performances were enhanced with the re-sampling method as the imbalanced dataset was identified to have performed poorly. Attribute Selection with the top ten attributes and 10-fold cross-validation offer best performances. The six predictive models utilized in this study are Linear Discriminant Analysis, Logistic Regression, Classification and Regression Tree, K-Nearest Neighbour, Naïve Bayes Classifier, and Support Vector Machines. Classification and Regression Tree model and Linear Regression had the best accuracy score of 0.86 after 10-fold cross-validation and top ten attribute selection. This study concludes that student behavioral attributes are useful predictors of student success.**

***Keywords*—Blended-Learning, Data Re-sampling, Machine Learning Models, Information Gain, Feature Selection, SMOTE**

I. Introduction

The machine learning technique is one of the main methods used in studying student performance or success, aside from statistical analysis and data mining. Academic performance is a daunting challenge for tertiary education institutions across the globe. [1] and [2] described data analytics as a tool for identifying students who are struggling educationally and enhancing throughput in various educational institutions.

This study falls under the category of Educational Data Mining (EDM). EDM is a subdivision of data mining that specializes in designing, evaluating, and implementing various automatic tools for measuring vast amounts of data from academic environments [3]. This study investigates student success through student behavioral attributes in the learning management environment. In any learning environment, student engagement is a crucial indicator for assessing a student's success or failure [4]. The channels of education delivery include traditional classroom, online-learning, blended-learning, and others. The Learning Management System (LMS) is a learning platform that allows instructors and learners to communicate without having to meet in person [5]. The global adoption of LMS platforms in learning is increasing by the day as several factors have warranted this acceptance. The reason for the adoption of LMS include the convenience of learning at student pace, improvement of cost-efficiency for the institutions and full coverage of a large number of students [6]. The blended-learning, which is interchangeably called hybrid-learning, is an infusion of both standard classroom and technology-aided settings [7]. [8] provided objectives of blended-learning as an effective learning process, student-teacher physical contact, academic performance enhancement, and learner's freedom. However, the reported rate of failure in blended-learning in the undergraduate programs has dramatically increased in current time [9]. Research into the determinant factors for a boost in student success in this blended-learning environment will increase throughput in tertiary education across the globe as the present COVID-19 pandemic necessitated the adoption of one form of online or the other. This research provides machine learning models with optimal performance in predicting student success in undergraduate as a form detective tool in assisting students from not dropping out of the blended-learning course and increasing academic outcomes.

In this paper, Section II discusses the related work on the prediction of the success of students with the application of machine learning. Section III highlights the research design, including data collection and pre-processing, feature selection, machine learning techniques, and evaluation metrics. Section IV outlines the results. Section V concludes the paper.

II. Related Work

Student success prediction is a crucial challenge in the technology-aided settings [5]. Past researches have provided various factors that influence student success. Some the factors are the student demographic information such as gender [10], previous academic performance [11] and interactions with the learning environment [6]. Therefore, this research investigates the influencing power of the student's interaction with the Learning Management System. Most scholars proposed that

978-1-6654-4067-7/21/$31.00 © 2021 IEEE

the engagement/interaction on LMS has a positive correlation with student success [6], [12]. [13] studied student performance on final examination grade in an undergraduate program using Decision Tree, Naïve Bayes, Logistic Regression, Support Vector Machine, K-Nearest Neighbour, Sequential Minimal Optimisation and Neural Network. The Logistic Regression performance was the best among the algorithms used in the study with an accuracy value of 66%. The authors in [14] predicted student academic performance in the virtual learning environment for four categories. Artificial Neural Network outperformed Logistic Regression and Support Vector Machine algorithm with a classification accuracy of 84% to 93%. [15] used the Recurrent Neural Network (RNN) in their indepth knowledge and engagement study. The authors achieved an accuracy value of 88.3% for RNN in their research. [5] investigated student learning performance from LMS data using Support Vector Machine, Linear Discriminant Analysis, Random Forest, K-Nearest Neighbour, and Classification and Regression Tree (CART). The result showed the performance of random forest as the best with an accuracy value of 90%. [5] further suggested the class imbalance problem's solution in future work. [16] applied the Logistic Regression predictive model in their study on variables (degree of engagement, degree of prestige, degree of visibility, student access amount, management system by subject, experience, age, and gender) and got an accuracy value of 87.53%. [12] investigated the academic success of students based on their learning management network activities. The predictive models used in the analysis were Artificial Neural Network, Decision Tree, and Naïve Bayes with bagging boosting and ensemble techniques. The highest accuracy value of 82% was obtained from the Decision Tree classifier.

Understanding data sources in prediction are essential as it lays the groundwork for future research that has yet to be pursued. It also cuts down computation times on feature extraction [17]. Many studies have used qualitative data (surveys), quantitative data (student behavioral data from online learning activities), and others used combinations of both data types in the prediction of student performance [4]. [6] in their research used publicly available data from the Open University of the United Kingdom to investigate the student's performance. Four online courses from the Moodle LMS logfile data of the undergraduate students at Tel Aviv University, Israel, were utilized by [18] used.

Captured images from videos were developed and used for engagement recognition in relation to student performance by [19]. [20] investigated the connection between student engagement and academic success in a technology-mediated platform using the LMS data from a North American university undergraduate science course. [21] explored interview questions for student engagement challenges in e-learning platforms at different Saudi universities and their relationship with student performance. [16] applied Moodle LMS data from graduate courses in public management course at the university in Brazil from 2014 to 2015. [6] used four variables to analyze student performance in online learning: initial evaluation results, the highest level of education, final test score, and clicks on the learning site. Academic achievement was strongly correlated with student clicks on nine online learning sites, final grades, and evaluation performance. Students click on forumng and oucontent were also found to be influential in predicting student performance. Student participation and final exam grade were positively impacted by forum conversation and access to course content. [20] examined the connection between student engagement and performance in a technology-mediated setting using nine engagement metrics and a cluster analysis derived through students' activity records. According to their findings, student characteristics such as frequency of logins, material read, and the number of forum read affected quiz results, resulting in a high final course score. Because of the positive association between participation and results, [20] suggested that student engagement may be a determinant of academic success.

III. Methodology

In this study, we seek to predict the success using the student behavioral patterns/activities on learning management. Six machine learning predictive models will be trained with K-fold cross-validation after feature selection using the top five and ten features after applying the information gain filter on the dataset obtained from LMS. The confusion matrix, accuracy, precision, recall, and f1-score will be used to evaluate the models' performance in this study to ascertain the model with the optimum performance.

A. Data Acquisition and Pre-processing

Students dataset containing the demographic, behavioral, and academic records were obtained from Kalboard 360 LMS. The data has 480 records of 305 male and 175 female students in an institution [12]. The data consists of 16 numerical and categorical attributes of students gathered over two semesters(first and second). The target variables are low, medium, and high represented below:

$$Grade = \begin{cases} 0-69, & \text{Low} \\ 70-89, & \text{Medium} \\ 90-100, & \text{High} \end{cases}$$

The dataset's class distribution is 127, 211, and 142 for low, medium, and high classes. To fix the class imbalanced distribution challenge, we will be employing Synthetic Minority Oversampling Technique (SMOTE) to avoid dominance from the majority class and improve the models' performance. The SMOTE works by generating new instances from the minority group prior to training the model [22].

B. Attribute Selection

To predict student success in a blended-learning environment, we explored the Information Gain evaluation filter to determine the most contributing attribute. For attribute selection, entropy is used to measure the value of attributes in descending order. We will be selecting the attributes with high entropy for dimensionality reduction and improvement in model performance [23]. Information gain value is represented mathematically as $0 \leq e \leq 1$. This means that the value spans from 0 to 1.

C. Classification Models

In this study, two linear (Linear Discriminant Analysis and Logistic Regression) and four non-linear (Classification and Regression Tree, K-Nearest Neighbor, Naïve Bayes and Support Vector Machines) supervised machine learning models will be trained to predict the success of the student in a blended-learning setting.

a) Linear Discriminant Analysis: The Linear Discriminant Analysis (LDA) is a predictive model used to model the differences in classes. The discrimination is done by comparing the means of attributes. LDA is used for dimensionality reduction and the minimization of the possibility of misclassifying cases. The structure of LDA used in this paper is from [24].

b) Logistic Regression: The Logistic Regression (LR) is used for predictive exploration. LR is also used to forecast categorical dependent attributes, mostly with the support of predictor attributes. The LR is focused on the estimate of the greatest probability, and the estimate must be most likely. The architecture of the LR used in this study follows [25].

c) Classification and Regression Tree: The Classification and Regression Trees (CART) build a framework from the training set. The division points are selected rapaciously by comparing each attribute and the significance of each attribute in the training set to reduce the loss function [24]. The application of the CART model in this paper is from [24].

d) K-Nearest Neighbor: The K-Nearest Neighbor (KNN) identifies the centroid sample in the training dataset for new samples. The overall average sample is considered here as forecast from the centroid closest neighbor [5]. The distance measure used is the Euclidean distance. KNN is a simple model but very useful in prediction.

e) The Naïve Bayes Classifier: The Naïve Bayes Classifier (NBC) is the most proactive and logical learning algorithm for most classification problems. NBC is based on Bayes' theory of strong assumptions of independence within attributes using a Bayesian framework [1], [6], [24]. The execution of NBC in this study is gotten from [24]

f) Support Vector Machines: The Support Vector Machines(SVM) is an efficient, strong, and reliable predictive model that is identified by a separating hyperplane. SVM generates a decision boundary that is used for prediction. SVM works by determining the closest data dimensions called support vectors to the inference segregation in the training dataset and separates the current test variable through the use of the functional margin [26]. The implementation of the SVM used in this paper comes from [26]

D. Performance Metric

The efficiency of the six classification models used in this study will be assessed through performance metrics such as the confusion matrix, accuracy, recall, precision, and F1 score after K-fold cross-validation, where k=10. These performance metrics have been widely used in previous research.

Table I: Information Gain and Attributes Categorisation for the Set of Attributes in the Prediction of Student Success

Rank	Entropy	Attribute	Attribute Categorisation
1	0.46	Visited Resources	Behavioural
2	0.4	Student Absence Days	
3	0.37	Raised Hands	
4	0.26	Announcement View	
5	0.15	Parent Answering Survey	
6	0.13	Nationality	Demographic
7	0.13	Relation	
8	0.12	Place of Birth	
9	0.11	Discussion	Behavioural
10	0.10	Parent School Satisfaction	
11	0.07	Topic	Academic
12	0.05	Gender	Demographic
13	0.04	Grade Id	Academic
14	0.01	Semester	
15	0.01	Stage Id	
16	0.00	Section Id	

- **Confusion Matrix**: The value of the information provided by a predictive model about expected and actual class labels is held in the confusion matrix. Our models are evaluated using the information in the matrix.
- **Accuracy**: The accuracy score is a common metric for evaluating classification models. It's calculated as the number of precise predictions dependent on the total number of predictions.
- **Recall**: The number of accurate positive predictions and the ratio of the total number of positives are referred to as recall. This is also known as the true positive rate.
- **Precision**: The number of accurate positive predictions as a percentage of the total number of positive predictions is known as precision.
- **F1 score**: The F1 score represents the average of recall and precision. It serves as a red flag of incorrectly classified performance.

IV. Results and Analysis

The results of the experiments performed with the six predictive models are presented here. After applying the information gain attribute evaluation in Table I, ten attributes were the most contributing attributes in predicting student success. The top five attributes are categorized under student behavioral attributes with the LMS, as shown in Table I. The two separate experiments were performed using the top 5 and 10 attributes for attribute selection to reduce the dimension and improve models' performances. K-fold (5 and 10) cross-validations were utilized for training the models after re-sampling the data with the SMOTE technique in this study.

The Information Gain results show the order of the importance of the features in the prediction of student success in descending order. Figure 1 also gives the graphical representation of the features in order of entropy value.

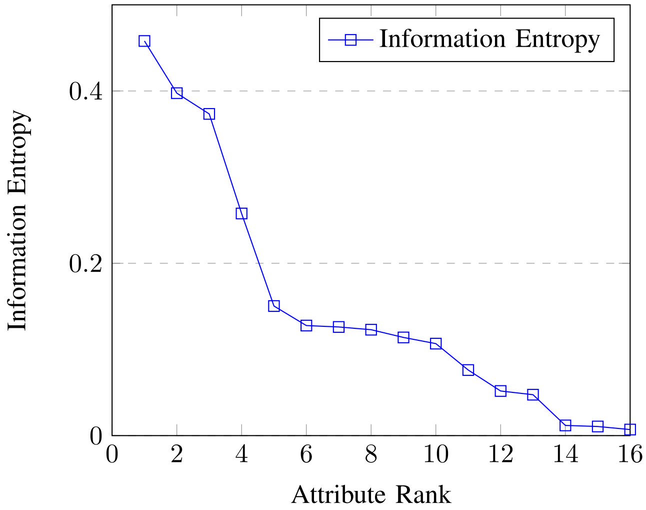

Figure 1: A Chart representation of the Information Gain for Set of Attributes to Predict Student Success

		Predicted		
		Low	Medium	High
Actual	Low	98%	2%	0%
	Medium	19%	62%	19%
	High	0%	10%	90%

Table II: Confusion Matrix showing the performance of **Linear Discriminant Analysis** Model after 10-fold cross-validation using the top 10 attributes.

		Predicted		
		Low	Medium	High
Actual	Low	98%	2%	0%
	Medium	14%	67%	19%
	High	0%	10%	90%

Table III: Confusion Matrix showing the performance of **Logistic Regression** Model after 10-fold cross-validation using the top 10 attributes

A. *Prediction Models*

This segment of the paper presents the results obtained from six models used in this study. The confusion matrices in Tables II, III, IV, V, VI and VII highlight the performances of the predictive models. The confusion matrices are determined from the testing dataset with the SMOTE-balanced dataset after the 10-fold cross-validation. We obtained the best model performances from the balanced dataset in contrast to the imbalanced dataset performance. Table VIII and Table IX also compared the performances of the application of the 10 and 5 fold cross-validation on the balanced and imbalanced dataset. The performance of the 10-fold cross-validation outweighs that of 5-fold in the balanced dataset, while 10 and 5 fold cross-validate make no difference in the models' performances in the imbalanced dataset. From Table II, the Linear Discriminant Analysis obtained an accuracy score of **0.84** after attribute selection of the top 10 attributes and 10-fold cross-validation. The percentage of the classification of the classes are 98, 62 and 90 for the low, medium, and high, respectively. The accuracy score for the Logistic Regression model is **0.86** and Table III shows the performance of Logistic Regression for low-class as 98%, medium as 67%, and high as 90%. Table IV represents the Classification and Regression Tree model's performance with an accuracy score of **0.86**. The percentage of the low, medium, and high classes is 95, 72, and 88, respectively. The K-Nearest Neighbour accuracy score is **0.81** which was obtained from the confusion matrix in Table V with the low, medium, and high classes percentage as 92, 56, and 90. The dominant class in this classification is low-class. The Naïve Bayes Classifier performance as represented in Table VI gives the accuracy score of **0.82** where the classes are rightly classified in the percentage of 88, 61, and 92 for low, medium, and high. The performance of Support Vector Machines as illustrated in Table VII presents an accuracy score of **0.72** achieved from the classification of 90%, 36%, and 84% for low, medium, and high classes. The best accuracy were achieved from the **Logistic Regression** and **Classification and Regression Tree** with an accuracy score of **0.86**. The poorest performance was obtained from the **Support Vector Machines** with an accuracy score of **0.72**. In summary, the performances of the six models used in this study are represented in Table X where the Classification and Regression Tree outperformed the five other models with an accuracy value of 0.86, precision value of 0.86, recall value of 0.86, F1-score value of 0.86 and Area Under Curve (AUC) value of 0.97. The other models in order of their performances are Logistic Regression, Linear Discriminant Analysis, Naïve Bayes Classifier, K-Nearest Neighbour, and Support Vector Machines.

V. Conclusion

This paper gives a framework for predicting student success in a blended-learning course to assist vulnerable students who are prone to fail or withdraw from the program. The result represented in Table I shows that the behavioral and demographic attributes are the most contributing predictors in forecasting student performance. The academic attributes have no tangible contribution to attribute importance. The

		Predicted		
		Low	Medium	High
Actual	Low	95%	5%	0%
	Medium	6%	72%	22%
	High	0%	12%	88%

Table IV: Confusion Matrix showing the performance of **Classification and Regression Tree** Model after 10-fold cross-validation using the top 10 attributes.

		Predicted		
		Low	Medium	High
Actual	Low	92%	5%	3%
	Medium	28%	56%	16%
	High	2%	8%	90%

Table V: Confusion Matrix showing the performance of **K-Nearest Neighbour** Model after 10-fold cross-validation using the top 10 attributes.

		Predicted		
		Low	Medium	High
Actual	Low	88%	12%	0%
	Medium	11%	61%	28%
	High	0%	8%	92%

Table VI: Confusion Matrix showing the performance of **Naïve Bayes** Model after 10-fold cross-validation using the top 10 attributes.

		Predicted		
		Low	Medium	High
Actual	Low	90%	8%	2%
	Medium	36%	36%	28%
	High	2%	14%	84%

Table VII: Confusion Matrix showing the performance of **Support Vector Machines** Model after **10-fold cross-validation** using the **top 10 attributes**

Predictive Model	**SMOTE + FS**		**Raw_data +FS**	
	5 Features	**10 Features**	**5 Features**	**10 Features**
LDA	0.82	0.84	**0.75**	**0.74**
LR	**0.84**	**0.86**	0.74	**0.74**
CART	0.82	**0.86**	0.69	0.73
KNN	0.75	0.81	0.58	0.60
NBC	0.82	0.82	0.71	0.67
SVM	0.70	0.72	0.56	0.60

Table VIII: Comparison of the Models' Performances in terms of the Accuracy score for balanced (SMOTE) and Imbalanced (raw) data after **10-Fold Cross-Validation** using 5 and 10 Features.

Predictive Model	**SMOTE + FS**		**Raw_data +FS**	
	5 Features	**10 Features**	**5 Features**	**10 Features**
LDA	0.81	0.83	**0.75**	0.74
LR	0.83	**0.84**	**0.74**	**0.74**
CART	**0.86**	**0.84**	0.69	0.73
KNN	0.74	0.76	0.58	0.60
NBC	0.83	0.82	0.71	0.67
SVM	0.70	0.72	0.56	0.60

Table IX: Comparison of the Models' Performances in terms of the Accuracy score for balanced (SMOTE) and Imbalanced (raw) data after **5-Fold Cross-Validation** using 5 and 10 Features.

Model	**Accuracy**	**Precision**	**Recall**	**F1 score**	**AUC**
LDA	0.84	0.84	0.84	0.83	0.94
LR	**0.86**	0.85	**0.86**	0.85	0.94
CART	**0.86**	**0.86**	**0.86**	**0.86**	**0.97**
KNN	0.81	0.81	0.81	0.80	0.92
NB	0.82	0.81	0.82	0.81	0.92
SVM	0.72	0.70	0.72	0.70	0.90

Table X: The Summary of the Performance Metrics of the Predictive Models' Performances after 10-fold Cross-Validation on the SMOTE balanced dataset and top 10 Attribute Selection.

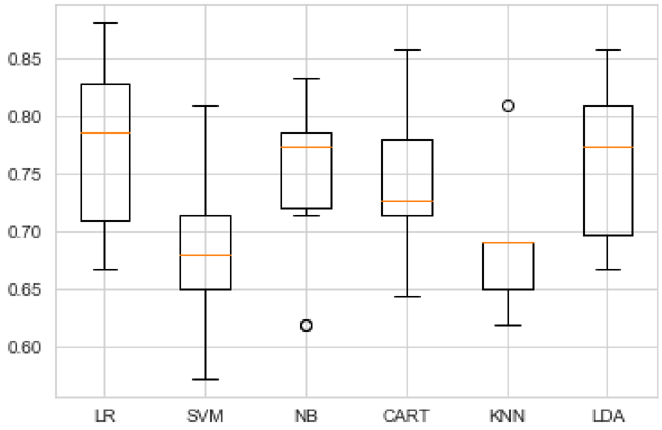

Figure 2: Algorithm Comparison Performance by Training Set

imbalanced data resulted in the models' poor performances in this study; hence, SMOTE method re-sampling technique for balancing to equal count. The application of SMOTE method and the attribute selection using the top 10 features recorded high-performances in the models' performances as expressed in Table VIII. The 5-fold cross-validation results in tab: 5-fold comparison were not as good as that of 10-fold cross-validation in Table VIII except for CART and NBC with the top 5 features. It is also important to note that the 5 or 10 fold cross-validation results are the same for the imbalanced data with 5 and 10 top attributes as shown in Table IX and Table VIII. In Figure 2, the LR and LDA performances were the best at the training phase. From Table VIII, the accuracy scores of CART and LR were the best, with a score of 0.86. The other accuracy scores are 0.84, 0.82, 0.81, and 0.72 for LDA, NBC, KNN, and SVM respectively. On critical analysis of the results illustrated in Table X, the CART performance for classification of student success band was the best with an accuracy score of 0.86, precision score of 0.86, recall of 0.86, and an AUC of 0.97. The other models in order of performances are LR, LDA, NBC, KNN, and SVM. We observed that the class's misclassification rate in medium-class was higher than any other classes for imbalanced and SMOTE balanced data. The most rightly predicted class is low, followed by high class.

In summary, the results obtained from this study show that machine learning techniques are efficient in identifying student performance on time for possible aid to prevent failure in their courses. The behavioral and demographic attributes are also essential in student performance classification. The constraint of this study is the type of data used. A more robust dataset in terms of the number of attributes would have given a holistic view of other essential attributes needed to classify the student performance. The findings of this study are solely dependent on the data utilized in the research. In the future study, we intend to investigate the reasons for the medium class's low classification for both imbalanced and balanced datasets. This study's machine learning models will also be extended to data from the university repository and not freely available dataset online used in this study. This paper's contribution is the presentation of the critical behavioral attributes for timely identification of students who are prone to withdraw from their courses for assistance by the institutions or administrators in an expeditious manner. This study concludes by highlighting that students' behavioral activities with the LMS are positive predictors to detect the students' performance.

Acknowledgment

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

References

[1] 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.

[2] 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.

[3] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert systems with applications*, vol. 33, no. 1, pp. 135–146, 2007.

[4] 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.

[5] A. Dutt and M. A. Ismail, "Can we predict student learning performance from lms data? a classification approach," in *3rd International Conference on Current Issues in Education (ICCIE 2018)*. Atlantis Press, 2019, pp. 24–29.

[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] 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.

[8] R. T. Osguthorpe and C. R. Graham, "Blended learning environments: Definitions and directions," *Quarterly review of distance education*, vol. 4, no. 3, pp. 227–33, 2003.

[9] 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.

[10] Z. Cai, X. Fan, and J. Du, "Gender and attitudes toward technology use: A meta-analysis," *Computers & Education*, vol. 105, pp. 1–13, 2017.

[11] C. J. Asarta and J. R. Schmidt, "Comparing student performance in blended and traditional courses: Does prior academic achievement matter?" *The Internet and Higher Education*, vol. 32, pp. 29–38, 2017.

[12] 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.

[13] A. S. Hashim, W. A. Awadh, and A. K. Hamoud, "Student performance prediction model based on supervised machine learning algorithms," in *IOP Conference Series: Materials Science and Engineering*, vol. 928, no. 3. IOP Publishing, 2020, p. 032019.

[14] H. Waheed, S.-U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, "Predicting academic performance of students from vle big data using deep learning models," *Computers in Human Behavior*, vol. 104, p. 106189, 2020.

[15] K. Mongkhonvanit, K. Kanopka, and D. Lang, "Deep knowledge tracing and engagement with moocs," in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 2019, pp. 340–342.

[16] J. C. S. Silva, J. L. Ramos, R. L. Rodrigues, A. S. Gomes, F. d. F. de Souza, and A. M. A. Maciel, "An edm approach to the analysis of students' engagement in online courses from constructs of the transactional distance," in *2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT)*. IEEE, 2016, pp. 230–231.

[17] J. Gardner and C. Brooks, "Student success prediction in moocs," *User Modeling and User-Adapted Interaction*, vol. 28, no. 2, pp. 127–203, 2018.

[18] T. Soffer and A. Cohen, "Students' engagement characteristics predict success and completion of online courses," *Journal of Computer Assisted Learning*, vol. 35, no. 3, pp. 378–389, 2019.

[19] A. Kamath, A. Biswas, and V. Balasubramanian, "A crowdsourced approach to student engagement recognition in e-learning environments," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–9.

[20] 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.

[21] M. A. Alsubhi, N. S. Ashaari, and T. S. M. T. Wook, "The challenge of increasing student engagement in e-learning platforms," in *2019 International Conference on Electrical Engineering and Informatics (ICEEI)*. IEEE, 2019, pp. 266–271.

[22] R. Longadge and S. Dongre, "Class imbalance problem in data mining review," *arXiv preprint arXiv:1305.1707*, 2013.

[23] R. Ajoodha, S. Dukhan, and A. Jadhav, "Data-driven student support for academic success by developing student skill profiles," in *2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*. IEEE, 2020, pp. 1–8.

[24] J. Brownlee, "Machine learning mastery with python," *Machine Learning Mastery Pty Ltd*, pp. 100–120, 2016.

[25] S. Sperandei, "Understanding logistic regression analysis," *Biochemia medica: Biochemia medica*, vol. 24, no. 1, pp. 12–18, 2014.

[26] G. P. S. Manu, "Classifying educational data using support vector machines: A supervised data mining technique," *Indian Journal of Science and Technology*, vol. 9, p. 34, 2016.