Proc. of the 3rd International Conference on Electrical, Communication and Computer Engineering (ICECCE)
12-13 June 2021, Kuala Lumpur, Malaysia

Importance of Data Re-Sampling and Dimensionality Reduction in Predicting Students' 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
Science Teaching and Learning Unit
Faculty of Science
The University of the Witwatersrand
Johannesburg, South Africa
kershree.padayachee@wits.ac.za

***Abstract*—In this paper, we present the importance of data pre-processing in predicting students' success. We implemented Principal Component Analysis for dimensionality reduction to achieve better model performance. Data re-sampling technique was also utilized to handle the imbalanced class problem that is one of the significant issues in effective classification in Educational Data Mining due to the nature of the data from educational settings. We also performed a comparative analysis on the impacts of Random Under-Sampling (RUS), Random Over-Sampling (ROS), and Synthetic Minority Over-Sampling Technique (SMOTE) to an imbalanced dataset used in this study. SMOTE and PCA techniques application offer better performance compared to RUS and ROS with PCA. Support Vector Machine had the best accuracy value of 0.94 after the application of SMOTE and PCA. The application of PCA on the imbalanced data also positively affected the accuracy of the models used in this study. We used other performance metrics to evaluate our models: Kappa, Area Under Curve, and Precision-Recall curve. Our finding shows that the predictive models can predict student success with the proper application of PCA and data re-sampling techniques.**

Keywords*—*Student performance, Random Under-Sampling (RUS), Random Over-Sampling (ROS), Synthetic Minority Over-sampling Technique (SMOTE) , Classification Models, Imbalanced Dataset, Evaluation Metric, Principal Component Analysis

I. Introduction

Tertiary education throughput and other learning outcomes are critical for any nation's growth [1]. The study of student performance or success in higher learning institutions has attracted lots of attention from various researchers in the past and recently. Education institutions have recorded massive data from Learning Management System (LMS) platforms [2] due to the extensive usage and the change in the education delivery due to the COVID-19 pandemic. It is essential to note that practically all the educational institutions at various levels across the globe have embraced one form of web-based learning or the other as a medium of education delivery due to necessity and a more extensive reach of all learners.

There are concerns about improving student learning outcomes, reducing drop out and increasing throughput propensities, especially in the technology-mediated environments, because of the importance to the nations, institutions, instructors, and other stakeholders. Lately, there is a broad perception that identifying and mitigating academic failure and preventive measures offer far more impact than remediation [3]. Some scholars have identified student engagement or involvement with the LMS as a significant indicator in determining the completion of their degree programs or performance in a course [4]. Student engagement is essentially the student's learning involvement or participation in a learning environment [4]. The student engagement measurement is a bit challenging in technology-mediated settings, unlike the conventional classroom setting where it is easier to detect student engagement or disengagement levels [4]. For this study's purpose, we seek to predict students' performance in a blended-learning course. Simply put,blended-learning combines regular classroom teaching with technology-assisted teaching [5]. The blended-learning combines the benefits of technology-aided and conventional classroom settings. Social interactions between students and teachers, better student participation, lower implementation costs than full-fledged online learning, uncomplicated accessibility to information, and enhanced student satisfaction and experience are just a few of the advantages of blended-learning [6]. In a survey conducted by [7] on online learning activities, it was discovered that 80% of all higher education institutions and 93% of doctoral institutions offer blended-learning courses. The online element of blended-learning is presented by [5] as the use of a network or technology to plan, create, pick, manage, and extend learning. From the study of [8], it was opined that the time allotted to online activities in a blended-learning course is proportional to the course performance. This paper applies the comprehensive analysis by comparing the predictive models used in this study to forecast the students' success early in a blended-learning course to prevent dropout and massive failure. The comparison study of PCA implementation and re-sampling technique is used in our prediction task. This study aims to determine the importance of the re-sampling methods and dimensionality reduction in predicting students' success

978-1-6654-3897-1/21/$31.00

in a blended-learning course.

The following is how the paper is designed: The related work is seen in Section II, and the methodology is discussed in Section III. The result is presented in Section IV. Finally, in Section V, the conclusions are highlighted.

II. Literature Review

The importance of studying student performance or success in higher-learning can not be over-emphasized because significantly reduced failure or dropout rate is a determinant to any nation's progress. There is increased research for prompt detection of failure tendency in students to avert failure and other non-performance effects [3]. In a comparative study of student success at a South African university by [9] to forecast students at risk of attrition, student high school final grades and bio-data information were analyzed using six classification algorithms. The accuracy evaluation metric employed in the study shows the performance of random forest as the best compared to other algorithms used with an accuracy score of 82%. The researchers in [**?**] concluded that effective detection and tracking of students with the possibilities of failure would result in higher academic achievement and more productive educational experiences. [10] predicted student performance in an LMS platform using the Learning Vector Quantization algorithm for feature selection on 16 student attributes using four predictive models, namely: random forest,k-nearest neighbors, linear discriminant analysis, and classification and regression tree. The authors recorded an excellent kappa statistic value of 0.85 for a random forest classifier. [10] also identified student behavioral attributes as contributing attributes in their prediction. The studies on the improvement of student academic performance were performed by [11] and [12] using three categorizations of student attributes (demographic, behavioral, and academic) in an LMS. The authors used three algorithms: neural network, decision tree, and naïve bayes classifiers in their studies. The results from [11] and [12] studies revealed that student behavioral attributes have a strong correlation with student performance. The authors obtained the best model performance from the decision tree classifier.

[4] study of student performance in relation to student engagement using various machine learning algorithms revealed that tree-based algorithms are suitable to forecast student performance. The authors in [4] also concluded that students' behavioral attributes with the LMS are vital in the student performance. In Educational Data Mining (EDM), one of the major concerns is the multi-class imbalanced problem because of variance in the distribution of students' academic performances. Research works in imbalanced datasets used numerous techniques of over and under re-sampling to balance the distribution. Several scholars have provided different views on the utility of the re-sampling methods [13]. There are multiple ways to solve an imbalanced problem in the class data to avoid bias and improve the model's optimality [14]. [15] research shows that SMOTE is preferred over Random Under-Sampling (RUS) and Random Over-Sampling (ROS) methods due to the potential loss of useful information when RUS eliminates majority class samples and ROS reuses minority class samples. [16] utilized two different resampling strategies (SMOTE and RUS) to balance the class distribution problem. The author identified great improvement in the three machine learning models' performances in his study with SMOTE and RUS.

III. Methodology

This paper seeks to determine the effect of various resampling techniques to resolve the class-imbalanced problem while using different machine learning models and dimensionality reduction to predict student success in a blended-learning environment.

A. *Data Acquisition*

The dataset was acquired from Kalboard 360 Learning Management System. The dataset consists of 480 student record instances with 16 input attributes, and a target variable (low, medium, and high classes) [11]. The attributes were further divided into Personal (demographic), Educational (academic), and Engagement (behavioral) attributes. The class distribution of the dataset is represented as follows: 127, 211, and 142 for Low, Medium, and High classes, respectively.

B. *Data Pre-processing*

Data pre-processing is a critical stage in data mining because unprocessed datasets are converted into a machine intelligible form, and this stage also enhances machine learning model performance [8]. In order to resolve the class distribution problem represented in III-A, we will be utilizing and comparing Random Under-Sampling (RUS), Random OVER-Sampling (ROS), and Synthetic Minority Oversampling Technique (SMOTE) techniques for improved classification models' accuracy and to prevent bias from the majority class.

a) ***Random Under-Sampling (RUS)****:* The Random Under-sampling process selects a minimal group of majority cases whilst keeping the total minority cases. The RUS method chooses instances from the majority class at random [16].The shortcoming of utilizing the random under-sampling technique on the majority class to resolve the imbalanced class problem is the loss of insightful facts from the dataset that can help model forecast [15].

b) ***Random OVER-Sampling (ROS)****:* ROS creates fresh instances from the minority groups by arbitrarily choosing and repeating class labels from the minority class [17]. Certain models overfit and the computational time is longer with the ROS application, which is a drawback of using ROS to replicate the minority class in solving the class imbalanced problem [15], [17].

c) ***Synthetic Minority Oversampling Technique (SMOTE)****:* SMOTE functions by populating an X_1 case of a minority group and determining the closest minority class neighbors. The simulated model would then be formed by selecting one of the k closest neighbors X_2 and connecting X_1 and X_2 to form a line in the feature's space. The synthetic samples would be made up of a curvilinear combination of

the X_1 and X_2 samples that have been chosen [15]. SMOTE does not lead to biased estimates because of the simulated sampling method it utilizes [14].

C. Dimensionality Reduction

Dimensionality reduction is the method of reducing the number of independent variables for a classification algorithm. A classification algorithm with a smaller feature space could perform better in predictions [18]. This study will utilize Principal Component Analysis to reduce attributes for a more streamlined and easier-to-understand illustration of target definition. Principal Component Analysis (PCA) is a heuristic and unsupervised algorithm for converting a large feature space to a smaller feature space by identifying relevant attributes containing the highest relevant information about the set of data. The features are chosen based on how much variability they introduce into the performance. The prime principal component is the attribute that produces the highest variability [18]. With the PCA, data analysis is feasible, and the computational time of the classification model is significantly reduced [19].

D. Classification Models

*a) **Support Vector Machines (SVM)**:* SVM is a predictive model that uses a collection of learning rules to analyze data and interpret classification problems. SVM is widely used in EDM due to its high overall accuracy [1]. SVM can effectively perform non-linear prediction using the kernel trick by transforming the inputs into large function spaces [18].

*b) **K- Nearest Neighbor(KNN)**:* The KNN algorithm finds the centroid sample in the training dataset for new instances. The average total sample is projected from the centroid closest neighbor in this case [10]. The Euclidean distance is the unit of measurement in. KNN is a simplified algorithm with a lot of predictive power.

*c) **Classification and Regression Tree (CART)**:* CART forms a structure from the training set. The preference points are decided rapaciously by analyzing each attribute and the importance of each attribute in training set to lessen the loss function [10], [18]. CART is also known as a decision tree, and it incorporates regression and classification trees in one algorithm [10]. The structure of the CART model in this study is from [10].

*d) **Gradient Boosting Tree (GBT)**:* GBT is a well-known machine learning method for its predictability and pace, mainly when dealing with large amounts of data. It lowers the chances of overfitting. It works by combining a learning algorithm with many poor learners who are concurrently linked to produce an effective learner [4]. It is a powerful method for developing forecasting models. GBT is applicable to a variety of risk functions and improves the accuracy of these functions over predictive models [20].

*e) **Multilayer Perceptron (MLP)**:* MLP is a feed-forward artificial neural network class. For instruction, it employs the supervised learning principle of back-propagation and requires at least three different layers: input, hidden, and output layers [21]. The input signal is collected for analysis by the input layer. The output layer performs the prediction and classification tasks, and the processing driver is given by the hidden layers located between the input and output layers [21]. The data in MLP flows from the input to the output layer in a forward direction, and the neurons learn using the back-propagation neural networks.

*f) **Linear Discriminant Analysis**:* LDA is a linear classification algorithm utilized to model the variances in classes. The distinction is achieved by matching the averages of attributes. LDA is used for feature space reduction and the minimization of the possibility of misclassifying instances [10].

E. Performance Metric

We will evaluate the performances of the classification algorithms used in this paper with different performance metrics after applying the k-fold cross-validation on training 70% of the set. For this study, K is equal to 10. The performance metrics to be employed are: **Confusion Matrix**, **Accuracy**, **Area Under Curve(AUC)**, **Kappa** and **Precision-Recall curve**. We will compute the confusion matrix on the 30% testing set.

IV. Results and Discussion

In this study, we conducted five experiments as shown from the results represented in Figure 7 and Tables I, II, III and IV. Before the data re-sampling on the imbalanced dataset, we utilized Principal Component 6 (PC6) for feature space reduction as shown in Figures 7. We chose PC6 using the "cumulative percent variance accounted for" criterion. The literature suggested the choice of cumulated variance of 70% to 80% [22].

Table I shows the results of the application of PCA on the imbalanced dataset and non-application of PCA on the imbalanced dataset. The findings show that SVM and KNN achieved an increase in their performances after the application of PCA. For instance, SVM increased by 14% while KNN moved by 9% after we implemented PCA. However, the performances CART, XGB, and LDA remained the same with or without PCA. The only model that benefited from the non-application of PCA in this phase of the experiment is MLP.

From the third experiment represented in Table II, we applied PCA to a SMOTE re-balanced dataset, and we did not apply PCA to the other SMOTE re-balanced dataset. The results revealed that six classification models(SVM, KNN, CART, GBT, MLP, and LDA) achieved an increase in accuracy values. SVM had the best performance with an accuracy value of 0.94, AUC 0.99, and Kappa value of 0.91. The finding in this phase of the experiment is that PCA influenced the models' performances.

The fourth experiment in Table III presents a rise in the performances of three of the models used in this study. The SVM, KNN, and MLP rose by 20, 7, and 15 with PCA implementation on the ROS re-sampled dataset. CART and

Table I: The Comparison of the Models' Accuracy, AUC, and Kappa on the Imbalanced Data after 10-Fold Cross-Validation with and without PCA.

Model	Imbalanced data + PCA			Imbalanced data		
	Accuracy	AUC	Kappa	Accuracy	AUC	Kappa
SVM	**0.77**	**0.91**	**0.63**	0.63	0.82	0.42
KNN	**0.69**	**0.85**	**0.53**	0.6	0.78	0.36
CART	0.7	0.9	0.51	0.7	0.9	0.51
GBT	0.72	0.87	0.55	0.72	0.87	0.55
MLP	0.69	0.86	0.53	0.75	0.9	0.61
LDA	0.75	0.89	0.61	0.75	0.89	0.61

LDA achieved a better performance without PCA, while GBT was the same with or without PCA application.

With the RUS re-balanced dataset in Table IV, we utilized PCA, and we did not use PCA on the other re-balanced data. This phase's results also show that PCA's application improved SVM, KNN, GBT, and LDA performances. The CART and MLP performed better without the implementation of PCA.

On the evaluation of the models with Precision-Recall curve, Figures 1, 2, 3, 4, 5 and 6 provide the graphical representation of precision against recall. A graph of precision versus recall for various probability levels is known as the precision-recall curve. An accurate predictive model is presented as a point at the location (1,1), and the graph tends towards the location (1,1)'s direction. The precision-recall curve's emphasis on the minority class indicates it as an essential assessment for imbalanced classification models. The literature suggested the evaluation with precision-recall curves for a very biased dataset [23].

The kappa value is a measurement that evaluates the actual and predicted accuracy, especially for an imbalanced class dataset. The kappa value is computed from the confusion matrix. A kappa of 100 percent indicates that the model agrees fully with the stochastic classifier, while a kappa of 0% indicates complete disparity [23]. For the models in this study, the highest kappa value was achieved by SVM as shown in Table II. The kappa values of all the models are closer to 100% as represented in Table II which makes the models' performances with the SMOTE and PCA excellent compared to other experiments.

The AUC value returns a holistic rating of a predictive model's performance. AUC shows how well a model can separate the classes from each other. The AUC value of 1 signifies a clear division of classes, and the mid-way (0.5) value indicates that the model randomizes the classes in prediction. Table II shows that the AUC values are close to 1 in SMOTE and PCA experiment compared to the results from other experimental phases in this study.

The confusion matrices for all the classification models are represented in V, VI, VII, VIII, IX, and X after 10-fold cross-validation on the SMOTE re-balanced data and PCA dimensionality reduction.

V. Conclusion

In this paper, we suggest using data re-sampling and dimensionality reduction in predicting student success in a blended-

Table II: The Comparison of the Models' Accuracy, AUC, and Kappa on the SMOTE after 10-Fold Cross-Validation with and without PCA

Model	SMOTE + PCA			SMOTE		
	Accuracy	AUC	Kappa	Accuracy	AUC	Kappa
SVM	**0.93**	**0.99**	**0.89**	0.71	0.83	0.56
KNN	**0.85**	**0.97**	**0.77**	0.79	0.92	0.69
CART	0.87	0.97	**0.81**	0.87	0.97	0.79
GBT	**0.86**	**0.96**	**0.78**	0.83	**0.94**	0.75
MLP	**0.9**	0.93	**0.84**	0.85	**0.94**	0.77
LDA	**0.83**	0.94	**0.75**	0.82	0.94	0.72

Table III: The Comparison of the Models' Accuracy, AUC, and Kappa on the ROS after 10-Fold Cross-Validation with and without PCA

Model	ROS + PCA			ROS		
	Accuracy	AUC	Kappa	Accuracy	AUC	Kappa
SVM	**0.75**	0.9	**0.6**	0.56	0.9	0.35
KNN	**0.7**	**0.87**	**0.55**	0.63	0.79	0.42
CART	0.64	0.82	0.46	**0.72**	**0.85**	**0.57**
GBT	**0.76**	0.9	**0.63**	0.74	0.9	0.6
MLP	**0.64**	**0.83**	**0.47**	0.49	0.75	0.32
LDA	0.74	0.89	0.6	0.74	0.89	0.6

Table IV: The Comparison of the Models' Accuracy, AUC, and Kappa on the RUS after 10-Fold Cross-Validation with and without PCA

Model	**RUS + PCA**			RUS		
	Accuracy	AUC	Kappa	Accuracy	AUC	Kappa
SVM	**0.73**	**0.89**	**0.57**	0.59	0.88	0.39
KNN	**0.65**	**0.85**	**0.48**	0.6	0.78	0.39
CART	0.68	0.81	0.52	0.68	0.81	0.52
GBT	**0.73**	0.89	**0.59**	0.71	0.89	0.56
MLP	0.68	0.84	0.53	**0.74**	**0.9**	**0.61**
LDA	**0.72**	**0.89**	**0.57**	0.67	0.87	0.51

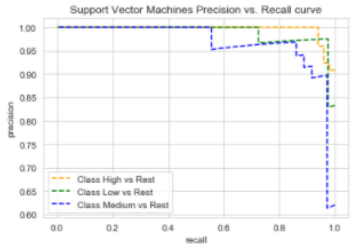

Figure 1: Support Vector Machines Precision-Recall curve

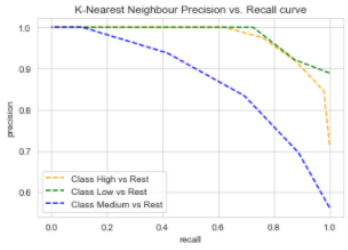

Figure 2: K-Nearest Neighbor Precision-Recall curve

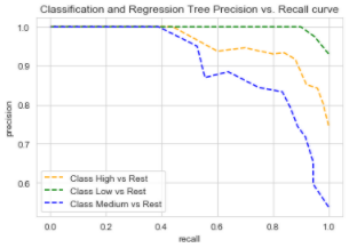

Figure 3: Classification and Regression Tree Precision-Recall curve

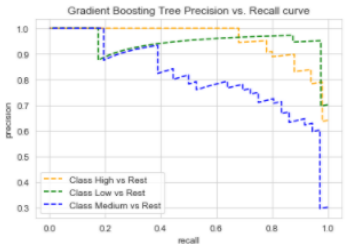

Figure 4: Gradient Boosting Tree Precision-Recall curve

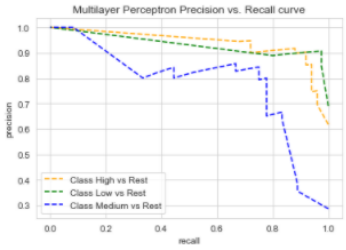

Figure 5: Multilayer Perceptron Precision-Recall curve

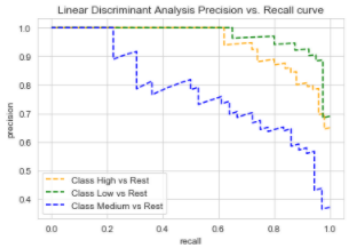

Figure 6: Linear Discriminant Analysis Precision-Recall curve

Table V: Support Vector Machines' Confusion Matrix after the application of PCA and SMOTE Re-sampling Technique.

		Predicted		
		Low	Medium	High
Actual	Low	95%	5%	0%
	Medium	3%	91%	6%
	High	0%	4%	96%

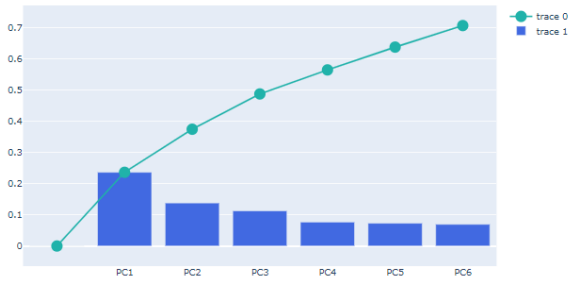

Figure 7: Principal Component Analysis Plot

Table VI: K-Nearest Neighbor's Confusion Matrix after the application of PCA and SMOTE Re-sampling Technique.

		Predicted		
		Low	Medium	High
Actual	Low	100%	0%	0%
	Medium	17%	69%	14%
	High	0%	10%	90%

Table VII: Classification and Regression Tree's Confusion Matrix after the application of PCA and SMOTE Re-sampling Technique.

		Predicted		
		Low	Medium	High
Actual	Low	100%	0%	0%
	Medium	6%	86%	8%
	High	0%	16%	84%

Table VIII: Gradient Boosting Tree's Confusion Matrix after the application of PCA and SMOTE Re-sampling Technique.

		Predicted		
		Low	Medium	High
Actual	Low	98%	2%	0%
	Medium	14%	69%	17%
	High	0%	12%	88%

Table IX: Multilayer Perceptron's Confusion Matrix after the application of PCA and SMOTE Re-sampling Technique.

		Predicted		
		Low	Medium	High
Actual	Low	98%	2%	0%
	Medium	11%	75%	14%
	High	0%	8%	92%

Table X: Linear Discriminant Analysis' Confusion Matrix after the application of PCA and SMOTE Re-sampling Technique.

		Predicted		
		Low	Medium	High
Actual	Low	97%	3%	0%
	Medium	17%	61%	22%
	High	0%	12%	88%

learning course. The impacts of these two data pre-processing techniques were examined exclusively on an LMS student dataset using six classification models (SVM, KNN, CART, GBT, MLP, and LDA).

The reduction of feature space on the imbalanced data provided an improved result on SMOTE, ROS and RUS balanced data. However, the dimensionality reduction did not improve the performance of the imbalanced (raw) data significantly. SMOTE re-sampling method with PCA achieved optimal performances in all the predictive models used in this study across the evaluation metrics used. The SMOTE balancing result affirms the claim in [15] regarding the choice of pre-processing sampling method for handling the class imbalance problem. In general, we realized that the results from re-sampling (SMOTE, ROS, and RUS) were better than the result obtained from the imbalanced class. The model with the best performance is the SVM after SMOTE and PCA applications. This study offers a platform for researches into other combinations of re-sampling techniques with feature space reduction. The result from this study presents an effective model for instructors, administrators, institutions, and other stakeholders' prompt intervention for students whose academic success are threatened by failure. Based on this study's findings and the data used, we conclude that data-resampling and dimensionality reduction are essential in the appropriate students' success forecast. In our future study, we plan to compare the re-sampling techniques utilized in this study with the latest techniques. We also intend to use a more robust dataset with a very high dimension and large-scale class variance.

Acknowledgment

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

References

[1] 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.

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

[3] C. Márquez-Vera, A. Cano, C. Romero, and S. Ventura, "Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data," *Applied intelligence*, vol. 38, no. 3, pp. 315–330, 2013.

[4] 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.

[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] R. M. Bernard, E. Borokhovski, R. F. Schmid, R. M. Tamim, and P. C. Abrami, "A meta-analysis of blended learning and technology use in higher education: From the general to the applied," *Journal of Computing in Higher Education*, vol. 26, no. 1, pp. 87–122, 2014.

[7] P. Arabasz, R. Boggs, M. Baker *et al.*, "Highlights of e-learning support practices," *Educause Center for Applied Research Bulletin*, vol. 9, pp. 1–11, 2003.

[8] C. Romero, M.-I. López, J.-M. Luna, and S. Ventura, "Predicting students' final performance from participation in on-line discussion forums," *Computers & Education*, vol. 68, pp. 458–472, 2013.

[9] R. Ajoodha, A. Jadhav, and S. Dukhan, "Forecasting learner attrition for student success at a south african university," in *Conference of the South African Institute of Computer Scientists and Information Technologists 2020*, 2020, pp. 19–28.

[10] 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.

[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] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Preprocessing and analyzing educational data set using x-api for improving student's performance," in *2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*. IEEE, 2015, pp. 1–5.

[13] N. V. Chawla, "Data mining for imbalanced datasets: An overview," *Data mining and knowledge discovery handbook*, pp. 875–886, 2009.

[14] S. T. Jishan, R. I. Rashu, N. Haque, and R. M. Rahman, "Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique," *Decision Analytics*, vol. 2, no. 1, pp. 1–25, 2015.

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

[16] Y. Chen, "Learning classifiers from imbalanced, only positive and unlabeled data sets," *Department of Computer Science, Iowa State University*, 2009.

[17] S. Huda, K. Liu, M. Abdelrazek, A. Ibrahim, S. Alyahya, H. Al-Dossari, and S. Ahmad, "An ensemble oversampling model for class imbalance problem in software defect prediction," *IEEE access*, vol. 6, pp. 24 184–24 195, 2018.

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

[19] S. Poudyal, M. Nagahi, M. Nagahisarchoghaei, and G. Ghanbari, "Machine learning techniques for determining students' academic performance: A sustainable development case for engineering education," in *2020 International Conference on Decision Aid Sciences and Application (DASA)*. IEEE, 2020, pp. 920–924.

[20] K. F. Hew, X. Hu, C. Qiao, and Y. Tang, "What predicts student satisfaction with moocs: A gradient boosting trees supervised machine learning and sentiment analysis approach," *Computers & Education*, vol. 145, p. 103724, 2020.

[21] 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.

[22] M. Richardson, "Principal component analysis," *URL: http://people. maths. ox. ac. uk/richardsonm/SignalProcPCA. pdf (last access: 3.5. 2013). Aleš Hladnik Dr., Ass. Prof., Chair of Information and Graphic Arts Technology, Faculty of Natural Sciences and Engineering, University of Ljubljana, Slovenia ales. hladnik@ ntf. uni-lj. si*, vol. 6, p. 16, 2009.

[23] L. A. Jeni, J. F. Cohn, and F. De La Torre, "Facing imbalanced data–recommendations for the use of performance metrics," in *2013 Humaine association conference on affective computing and intelligent interaction*. IEEE, 2013, pp. 245–251.