ISSN: 0974-8571 Vol. 13 No. 2 December, 2021

International Journal of Computational Intelligence in Control

Student Performance Prediction in Intelligent E-Learning for Tertiary Education

Mustafa Hameed and Nadeem Akhtar
Faculty of Computing, The Islamia University of Bahawalpur Pakistan,
gmhameed@yahoo.com, nadeem.akhtar.phd@gmail.com

Date of Submission: 25th August 2021 Revised: 29th October 2021 Accepted: 20th January 2022

How to Cite: **Mustafa Hameed and Nadeem Akhtar (2021). Student Performance Prediction in Intelligent E-Learning for Tertiary Education. International Journal of Computational Intelligence in Control 13(2).**

Abstract **- Tertiary education plays an important role in today's evolving environment. After COVID-19 pandemic, the pervasive availability of Learning Management Systems (LMS) has resulted in online educational activities with no geographic limitations or time constraints. Evaluation grades of online and on campus activities as grade book along with interaction patterns of students and teachers as event logs are recorded by LMS. There has been rapid growth in research on the prediction of student success or risk of abandonment with the help of advancements in data mining and machine learning techniques. This study predicts student performance at real time, by employing machine learning based learning analytics at tertiary education level. student performance prediction. We believe that the proposed research work provides teachers with a snapshot of student achievement and increases their interest in blended learning environments.**

INTRODUCTION

There are several areas where educational institutions may adopt technologies that will support educational leadership, teachers, and students.Artificial intelligence is one of the most innovative technologies which is widely used to support institutes to create rich teaching learning environment.Educational institutions are integrated into Education Management Information Systems (EMIS) to support teaching and learning. EMIS had evolved as learning management systems (LMS) in due course. However, the use of LMS-based e-learning has increased significantly everywhere since the COVID-19 pandemic.Interaction of all stakeholders, including students, teachers, and education administrators, is cost-effective with the assistance of the LMS. Geographical limits and time constraints are no more an issue. Alongside all these advantages, because of less or no face-to-face interaction, students can develop boredom with online learning systems over time. Also, for the teachers it becomes difficult to monitor and analyze student behavior towards learning. Educational institutions are integrating LMS with power of artificial intelligence to support and enrich teaching and learning process.

In general, tertiary education (TE) means post-secondary, technical, or vocational education carried out in colleges and universities. In this study TE refers to all sorts of education carried out in higher education institutions. Institutions combine on-campus educational activities with E-Learning platforms. This environment is referred as blended learning. However, most of the institutions offering distance or open educations solely relying on E-Learning platforms.

Primary goal of the learning analytics (LA) is to understand and improve entire learning process of students [1]. All the academic activities carried out online generate huge volume of data to store files and event logs of activities. LA identify learning patterns and predict performance of students using the raw data from LMS. This is achieved with statistical analysis and predictive modelling.

RELATED STUDIES

This section discusses the literature review and studiesrelated to our research. At the first, recent studies of learning analytics (LA) are discussed. Secondly, educational data mining and its differences with LA are discussed.

Learning analytics (LA) systems enable institutes to track student performance, achievements, and progress in near-real time, notifying any potential issues to instructors or support staff. They can then receive the earliest

possible alerts of students at risk of dropping out or under-achieving grades. Recent studies[2];[3]; and[4] show increasing interest of practitioner of educational management and research community in learning analytics.Learning analytics increasingly draw data from across the institution into a single learning records warehouse. This might include usage data from the learning management system (LMS) specifically attendance records and grade data. Backend of learning analytics is supported by machine learning algorithms. These algorithms are trained based on historical records of students. Evaluation strategies for learning analytics are proposed in[5]. Researchers attribute learning analytics as of a bricolage combining concepts from business intelligence system used in enterprises, academic analytics, educational data mining and ontologies from semantic web. Based on cognitive, social and teaching aspects, 13 different dataset are derived and analyzed by machine learning techniques in [6]. All datasets are derived from Moodle database, but researchers have noted that, there are no statistically significant differences among models.

Researchers in [1] has discussed different LA models and proposed an interesting models. There are research-based recommendations [7] that implementation of learning analytics systems at government level would support cross institute information about students. In another recent study [8], a Moodle plugin based on machine learning is developed. Keeping in view small datasets, researchers [9] identified ML classification models those are suitable to identify at-risk students from a small dataset. This study also identified that, early warning to at risk students in the start of the course is more suitable. In another relevant study [10] clustering and visualization of small datasets is performed. Support vector machine and learning discriminant analysis algorithms are found acceptable based on classification accuracy and reliability test rate. Data of a small cohort of students is analyzed in [11] and even with few attributes related to interim assessments, lecture attendances and records of interaction VLE. Results are promising and provide a valuable support for intervention to at risk students.

Performance of different machine learning techniques to predict student's dropout in a undergraduate course is compared in [12]. Study shows that performance of different algorithms varies from 77 to 93 % on unseen data. Researchers in [13] has analyzed clickstream data along with all activities carried out through VLE and have identified factors that impact the learning behavior of students while learning online.

By analyzing Open University Learning Analytics Dataset (OULAD) dataset, researchers in [32] compare results from deep learning and re-gression techniques. They compare results of whole dataset as well as demographic information, assessment and VLE interaction subsets of dataset. In [33] researchers have analyzed student's previous grades in university records and student responses related to demographic information. This analysis is performed by decision tree algorithms J48, REP-Tree and Hoeffding. Results show that J48 yields more accuracy that other two algorithms.

In context of distance education, researchers in [14] has analyzed student interaction data recorded by e-learning platform and derivative features.

Identification of at-risk students is performed by constructing two models, 1) at-risk student model and 2) learning achievement model by using machine learning techniques [15]. Performance of random forest, generalized linear model, gradient boosting machine and neural networks is compared. A novel recurrent neural network (RNN)-gated recurrent unit (GRU) joint neural network is proposed in [16] to predict student performance. This research study used statice personal biographic information and sequential behavior data with virtual learning environment. Student's individual characteristics and online learning behavior is analyzed in [17]. Combined feature set is produced based on time window constraint strategy and learning time threshold constraint strategy.

METHODOLOGY

This study for the construction of the student performance forecast model is divided into three components.In the first component, we pre-process the student's performance in the exam data set to remove missing and outlier values.Another second component selects a subset of suitable characteristics to be used in the design construction.Finally, machine learning classification algorithms are used to construct predictive models to predict student success based on comprehensive data and subsets of data.The results of the comparison should provide an appropriate subset of attributes for machine learning classifiers in this area.

In this section, existing methods and algorithms were evaluated and compared empirically in the context of the challenge of predicting examination performance.The classification can be binary or multinomial, according to the type of output variable (e.g.how many possible values a display variable can take). In this paper, a binary classification was applied as the problem was to predict whether the student is normal or at-Risk. In other words, determine which students were at risk of failing the class as opposed to those who did not.Thus, only two classes have been defined.In the case of a classification assignment, the goal was to categorize students into two categories, i.e. either "Normal" or "at-Risk".

Data preprocessing and feature selection are applied to a given set of raw data. Firstly, all missing values are removed. One of the pertinent issues in data quality is a missing value, which occurs when data values are not stored in the feature. The absence of data values can have a significant impact on sample representativity and statistical power. Hence, the estimation of these values is important for building accurate prediction models.

The proposed experimental design flow chart is illustrated in Figure 1, of the remainder of this section describing these components in detail.

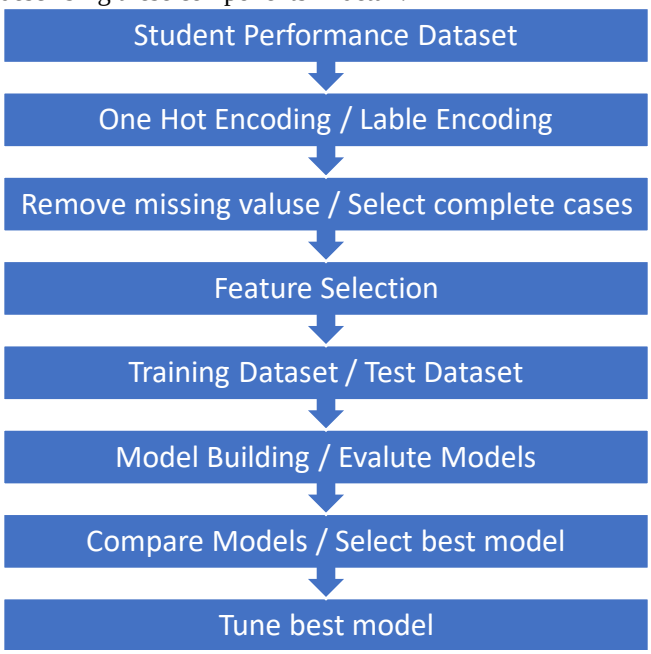

Figure 1 Workflow of proposed methodology

Table 1: Frequency of student status in dataset

Student Performance	Frequency	Percentage
Normal	409	51.06117
At-Risk	392	48.93883

Table 1: Frequency of student status in dataset summarizes the number of normal students and students at-Risk in dataset.Frequency of values depictsthe dataset almost balanced.

Table 2Attributes in dataset: a) original dataset, b) transformed dataset, c) dataset after feature reduction

(a)
gender
race/ethnicity
parental level of education
lunch
test preparation course
math score
reading score
writing score
(b)
gender
race.ethnicitygroup.B
race.ethnicitygroup.C
race.ethnicitygroup.D
race.ethnicitygroup.E
parental.level.of.educationbachelor.s.degree
parental.level.of.educationhigh.school
parental.level.of.educationmaster.s.degree
parental.level.of.educationsome.college
parental.level.of.educationsome.high.school
lunchstandard
test.preparation.coursenone
writing.score_new
reading.score_new
math.score_new
(c)
gender
lunchstandard
test.preparation.coursenone
writing.score_new
reading.score_new
reading.score_new

Feature selection (FS) means taking a specific data set and selecting the most useful and applicable features from it. If we have a dataset with d input features, feature selection will create a set of k features in such a way that k < d, with k being the smallest possible collection of relevant and important features. Original dataset consists of 8 attributes and after encoding dataset expanded to 15 attributes.Figure 6representfeature importance.

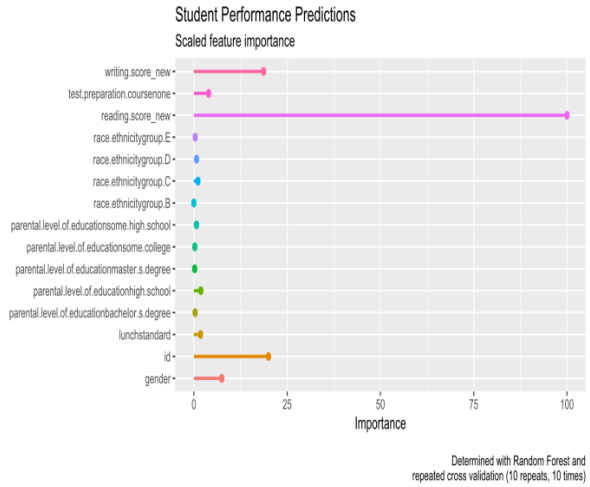

Figure 2 Feature importance

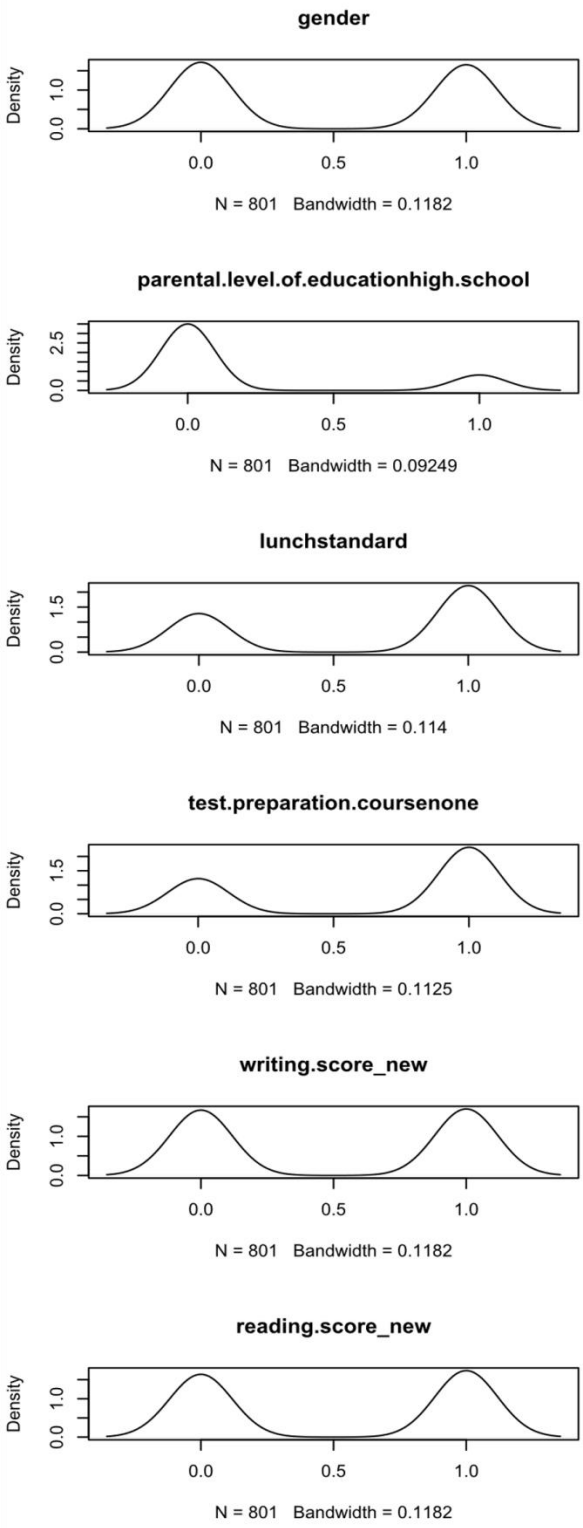

Figure 3 Density plot for reduced features

This study compares the performance of different numbers of classifiers in the context of student examination performance. The modelling process involves the selection of models based on various machine learning techniques used in the experiment. In this case various predictive models were used such as those based on decision tree, Bayesian method, logistic regression and SVM. The objective is to find the best classifier for the analyzed problem. Each classifier must therefore be trained on the featured set and the classifier with the best classification results is used for prediction. The classification algorithms taken into consideration are:

- Logistic Regression (LG),
- Linear Discriminate Analysis (LDA)
- Regularized Logistic Regression (GLMNET)
- k-Nearest Neighbors (KNN),
- Classification and Regression Trees (CART),
- Naive Bayes (NB)
- Support Vector Machines with Radial Basis Functions (SVM)

To evaluate the prediction models, we split the data into 70% for the training set and 30% for the evaluation set.The introduction of the validation package allows to preview the performance of the test package, causing an immediate revision of the model in case of unsatisfactory performance.In model training phase we used 10-fold cross validation with 3 repeats.

Accuracy and Kappa are the default measurements used to evaluate algorithms on binary and multi-class classification datasets in caret.Accuracy is the percentage of cases correctly ranked across all cases.It is more useful on a binary classification than multi-class classification problem because it may be less clear exactly how the precision decomposes between these classes.

Dataset is processed with the help of R language (version 4.0.2) in RStudio (version 1.3). Prediction results and learning process animations are represented using Shiny App. Shiny Apps are interactive web appli-cations built in R based on shiny package. Minimal effort is required to design and build responsive and powerful applications.Majority of the tasks for this study are performed in in RStudio for this study. Beside RStudio, MS excel for creating pivot tables, managing, and storing csv files was also used.

RESULTS

This section is dedicated to discussion on results for employed methodology.Although accuracy of all seven machine learning models is good at original dataset. But after going through our methodology accuracy have been improved significantly.

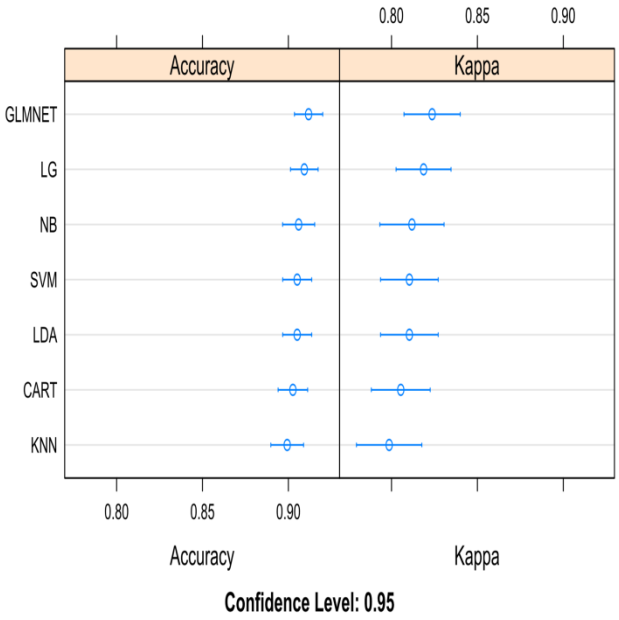

Figure 4Accuracy achieved by selected ML models

Table 3: Improvement in accuracy after employing proposed methodology

ML Model	Accuracy		
	Original Dataset	Transformed Dataset-2	Improvement
LG	81.73	91.09	9.37
LDA	81.31	90.51	9.20
GLMNET	81.98	91.09	9.12
KNN	80.73	91.14	10.40
CART	81.06	90.26	9.20
NB	81.40	90.93	9.53
SVM	81.31	90.80	9.49

Figure 5 represents the measure of accuracy of used machine learning models.Data transformation and feature reduction has resulted in improvement in accuracy of each machine learning model. But improvement in performance of KNN is significant.

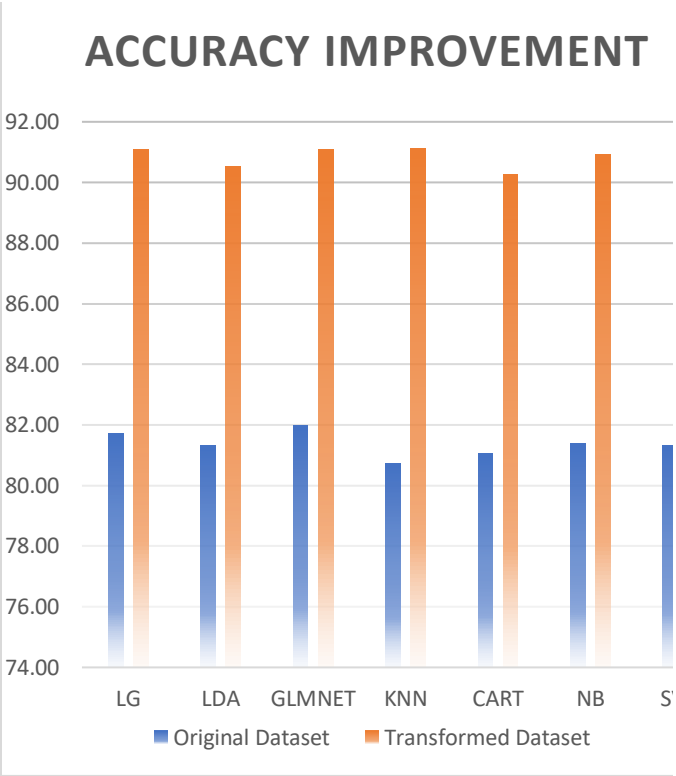

Figure 5Comparison of accuracy improvement

Accuracy comparison of ML models is shown in Figure 4. Accuracy of Logistic Regression (LG) model was 81.73 before transformation. After proposed methodology accuracy was 91.09 and improved9.37. Accuracy of Linear Discriminate Analysis (LDA) was81.31, after transformation 90.51 and improved 9.2.Regularized Logistic Regression (GLMNET)'s accuracy was81.98 which reach at 91.09 by improving 9.12. The best performing model k-Nearest Neighbors (KNN) initially gained accuracy of 80.73, but after transforming the dataset gone to 91.14 by improving10.4. Classification and Regression Trees (CART)'s accuracy was81.06 improving to 90.26 by a margin of 9.2. Naive Bayes (NB) performed initially at 81.4 and going to 90.93 improving to 9.53. Support Vector Machines with Radial Basis Functions (SVM) was at initially 81.31 and improved to 90.8 by the difference of 9.49.

Figure 6represents performance of four different ensemble methods. These four models include Bagged CART (BAG), Random Forest (RF), Stochastic Gradient Boosting (GBM) and C5.0 (C50).From this group of models, GBM outperformed all others. After comparing results, we can see that, overall k-Nearest Neighbors (KNN) was the best performing model for our transformed dataset.

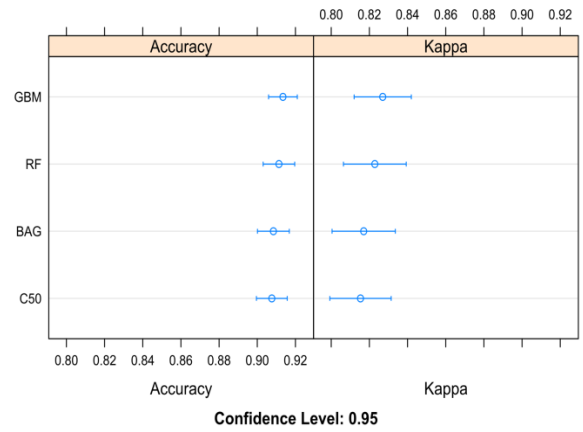

Figure 6 Performance of Ensemble methods on dataset

Figure 7 Shows the performance of KNN after fine tuning its parameters.

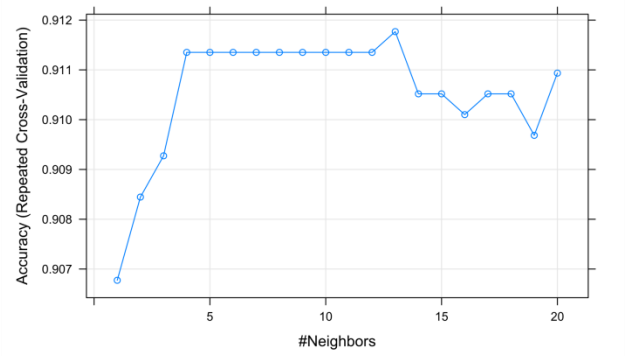

Figure 7 Fine Tuned KNN results

CONCLUSION

We applied machine learning techniques to identify the issues that can contribute to student performance prediction and, most importantly, to predict at-risk students. First, we assess statistically the data and then we classified them.The dataset was processed, divided between the training phase and the test phase, guaranteeing the same distribution of the target.We selected different classification algorithms and, for each one, we achieved the training and validation phases.From these, it was possible to compute the basic metrics required for an overall assessment (precision, recall, precision, f1 score, ROC curve, AUC, etc.)and identify the most appropriate classifier to predict at-risk students.The algorithm that gave the best results for the data set used was the KNN: it revealed the best accuracy rate (91.14).The results of the proposed automatic predictor demonstrate that the main attributes of performance prediction are gender, lunch, test

preparation, reading score, writing score and parental education.The results of the data analysis establish a starting point for the construction of increasingly effective student performance classifiers.The availability of additional data sets and the increase in the volume and attributes of the existing data set may reveal further dimensions.In addition, it is also necessary to guide students with recommendations to enhance their performance.

REFERENCE

1. Sciarrone, F. and M. Temperini. Learning Analytics Models: A Brief Review. in 2019 23rd International Conference Information Visualisation (IV). 2019. IEEE.
2. Aulck, L., et al. Predicting Student Dropout in Higher Education. in Workshop on #Data4Good: Machine Learning in Social Good Applications. 2016. New York, NY, USA: ICML.
3. Dewan, M.A.A., et al. Predicting dropout-prone students in e-learning education system. in 12th International Conference on Ubiquitous Intelligence and Computing. 2016. IEEE.
4. Rovira, S., E. Puertas, and I. Laura, Data-driven system to predict academic grades and dropout. PLoS ONE, 2017. 12(2): p. 1-21.
5. Dawson, S., et al. From prediction to impact: Evaluation of a learning analytics retention program. in The Seventh International Learning Analytics & Knowledge Conference on - LAK '17. 2017. Vancouver, BC, Canada: SOLAR.
6. Buschetto Macarini, L.A., et al., Predicting Students Success in Blended Learning—Evaluating Different Interactions Inside Learning Management Systems. Applied Sciences, 2019. 9(24): p. 5523.
7. Dawson, S., D. Gasevic, and T. Rogers, Student retention and learning analytics: A snapshot of Australian practices and a framework for advancement. 2016, Australian Government Office for Learning: Canberra.
8. Sáiz-Manzanares, M.C., R. Marticorena-Sánchez, and C.I. García-Osorio, Monitoring Students at the University: Design and Application of a Moodle Plugin. Applied Sciences, 2020. 10(10): p. 3469.
9. Leite, D., et al. Early detection of students at risk of failure from a small dataset. in 2021 International Conference on Advanced Learning Technologies (ICALT). 2021. IEEE.
10. Abu Zohair, L.M., Prediction of Student's performance by modelling small dataset size. International Journal of Educational Technology in Higher Education, 2019. 16(1): p. 27.
11. Wakelam, E., et al., The potential for student performance prediction in small cohorts with minimal available attributes. British Journal of Educational Technology, 2020. 51(2): p. 347-370.
12. Kabathova, J. and M. Drlik, Towards Predicting Student's Dropout in University Courses Using Different Machine Learning Techniques. Applied Sciences, 2021. 11(7): p. 3130.
13. Moreno-Marcos, P.M., et al., Analysis of the factors influencing learners' performance prediction with learning analytics. IEEE Access, 2020. 8: p. 5264-5282.
14. Queiroga, E.M., et al., A Learning Analytics Approach to Identify Students at Risk of Dropout: A Case Study with a Technical Distance Education Course. Applied Sciences, 2020. 10(11): p. 3998.