A Machine Learning Model to Predict Student Performance Rate

C UTHEJ[1], LOKESH C K[2]
[1, 2] *School of CSA, REVA University, Bangalore*

Abstract—The amount of perceived program success is ranked by the students who commit to the fulfillment of expectations for the program. It levels the rank satisfaction and importance of the primary usage program offered. The students acknowledge the approaches to measuring student success which is expected to pass/fail the program. Here it is recognized with the virtual learning environment interactions with the course information followed by the assessment and the marks given with respect to the tutor-given marks (TMA) and computer-graded marks (CMA) with the final exam. Here the model is to predict learning failures and the withdrawal of a student from the module presentation and some of the feature engineering has been done along with recommendations for further features that could be useful for this project. Few classification and regression models are employed to forecast student academic performance. This would give a student performance rate and give the failure rate to find out the risk. The best model for the regression task was Gradient with cross-validation of 4 folds and the best model for the classification task was the support vector machine classifier with a seventy-eight percent accuracy score.

Index Terms—primary usage, virtual learning environment, TMA, CMA.

I. INTRODUCTION

Student acknowledgment with the approaches to measuring student success rate which is expected to a student failure rate of the program. Here it is recognized with the virtual learning environment interactions with the course information followed by the assessment and the marks given with respect to the tutor given marks and computer-graded marks with the final exam. On the learning analytics on which a given dataset is used for this machine learning study the dataset, which is open to the public, includes tables with data on student demographics, modules taken when the modules begin (module presentations), and information on student's academic success as measured by grades for assignments and exams as well as interactions with the university's virtual learning environment.

The higher education sector has been using new technologies to collect more student data. This data is used to support the transfer of courses to the Internet. It has also been studied to see how it can impact learning. There have been over 200 scientific studies that have looked at the impact of this data. This shows the importance of sharing data in an open way so that we can learn from it and improve our education system. Open University is a large, distance-learning university. Teaching materials and other content are delivered to students via the VLE. Students' interactions with the educational materials are recorded and stored in the university's data warehouse. Each course is divided into small, individual parts that are repeated throughout the year. This way, you can always be sure to find the right part of the course to learn, no matter when it happens. Each part is named after the year and month it starts. For example, in October 2013, the part starting in January is called "A", and the part starting in February is called "B". The university offers a variety of courses, each of which can be studied separately or as part of a university program. You don't need any previous qualifications to enroll in these courses. Open University wants to make sure that your data is used in a way that is beneficial to you. They explain the policy on using student data and tell you that we will use your data for academic research purposes. They also make sure that your data will not be released to anyone who might be able to identify you. However, because the data in OULAD is anonymized, it is not possible to link it with individual students. Predicting which students will drop out or fail their modules and which will pass

them is the current task at hand. The dataset is way disorganized, with numerous missing values and some discrepancies between tables. The first section of the notebook has a complete list of cleaning procedures. Numerous discrepancies are brought to light and corrected, or recommendations are offered for how to do so in future work. Some feature engineering has been done along with recommendations for further features that could be useful for this project. Finally, a few classification and regression models are employed to forecast the academic performance of students.

II. LITERATURE REVIEW

The task is to predict the student's academic failure and student's withdrawal from module presentations. The final exam results are missing for all students except one of the modules. This means that using the entire table containing the scores as the regression response may lead to unreliable results because the information is not complete. This means that the student may pass the assignment, fail the final exam, and fail the entire module. Another point is that the score is the same as the result. So predicting the likelihood of failing when you know that her final grade is below 40% is not predictive at all. And it will be very interesting to see if we can identify students who are at risk of dropping out or failing without knowing anything about their actual academic performance.

Student performance is an important part of a university because one of the criteria for a quality university is based on excellent academic performance. In this study, a semi-supervised learning approach is used to rank the performance of undergraduate mathematics students at the University of Indonesia. A student's grades are divided into two categories, intermediate and advanced. In the clustering process, we use k-means clustering to divide the data into three clusters and choose the Naive Bayes classifier to classify them. The performance of the proposed algorithm is indicated by the accuracy, sensitivity, and specificity values, based on the results of this study, K-Means Clustering and Naive Bayes Classifier can be used to classify student performance.

Education is one concern of the Universities they focus on aim and provide an overview of the current state of research on Predictive Analytics in Higher Education. Highlight the most relevant instances of the predictor variables. They have provided in relation to EWS acting as a teacher support tool. Offering interventions to struggling students is limited. A major advancement in EWS is the ability to recommend ways to help potential students. Either way is the most effective. This may provide valuable information Implement effective interventions.

A predictive model of student performance in the early stages of Mixed Learning Course with Deep Neural Networks (NN) Architecture and use of online activity attributes as an input pattern. An experiment was then conducted is run to test model performance and predict the outcome (pass or fail) for both students. Results show that there are only a few Predictive performances that can be achieved early, specifically his first month of the course. In both accuracy and ROC_AUC values the more data collected by the third month, the better. The Highest Accuracy Achieved for Predicting the Final Results was good.

The relevant EDM research, including datasets and techniques used in the Research and identify the most effective techniques. many Key applications include predicting student performance, identifying undesirable student behavior, grouping students, and student modeling. These applications are designed to help decision-makers In educational institutions understand the situation of students, Improve student performance and identify learning priorities Develop different groups of students and learning processes. of Selected as a criterion for predictive accuracy Effectiveness of educational data mining techniques. Effective Techniques for Predicting Student Performance, Social Network analysis is the best technique for detecting unwanted data Student behavior and clustering and social network analysis The Most Effective Techniques for Grouping Students Modelling or this research suggest doing more A Comprehensive and Expanded Study to Assess Efficacy EDM technology with extended evaluation criteria.

The data analysis techniques are aimed at extracting hidden knowledge from data. The data mining techniques used in the research are Cluster analysis and decision trees. Cluster analysis was performed by organizing the collection of patterns into groups It is based on the similarity of student behavior when using course materials. The emergence of big data stored in databases containing records of student behavior in e-courses in higher education institutions. Since education data mining is part of the data mining field, cluster analysis and decision tree techniques were applied. A decision tree method was applied to the grouping results to enable a more detailed analysis of student behavior in the teaching and learning process.

As education plays an important role not only in an individual's life but also in the nation as a whole. Many students drop out of various academic courses each year. This study gave the contribution of student demographics to academic performance. A random forest classification model is used to predict a student's final exam grade. They have evaluated the attributes and their impact on student outcomes using three public datasets with different demographics. Holdout and cross-validation methods are used to evaluate experimental results. A random forest with three different datasets was yielded and this study gave relevance for educational authorities to predict student performance before they drop out. This can help institutions predict student work early on and pay special attention to weaker students from the very beginning of the session.

The paper attempts to provide models with optimal accuracy to identify students who need support to improve their academic ability and other learning outcomes. The examined effect Resampling of SMOTE data and the impact of attribute selection of this study. An imbalanced dataset was detected, so the model's performance was improved with a resampling method It cuts badly. Attribute selection with the 10 best attributes and 10-fold cross-validation provide the best performance. The predictive model used in this study is linear discriminant analysis, logistic regression, classification, and regression trees, K Nearest Neighbor, Naive Bayes Classifiers, and Support Vectors machine. Classification and Regression Tree Models and Linear The regression showed the highest accuracy score of 0.86 after 10-fold cross-validation and top 10 attribute selection.

Predicting student performance in a variety of ways to relevant information has evolved into an efficient tool for educational institutions to improve their curricula and teaching methods. Automated analysis of educational data Uses state-of-the-art machine learning techniques (ML) and artificial intelligence Intelligent algorithms (AI) are an active research area. This paper addresses student performance issues Prediction using three ML algorithms like Support Vector Classifier (SVC), k nearest neighbor (k-NN), and artificial neural functions Network (ANN) in the Open University dataset. Beneficial Data are analyzed including three main indicators demographics, engagement, and performance. The experimental analysis found the k-NN approach to be the best OU experiment in the comparison of applied and applied existing literature attributed to improved results changes to missing value handling and standardization of data approach.

The offered courses in a blended eLearning mode with a total of 6 sessions available Each course consists of 3 face-to-face lessons and 3 online lessons. Pure online courses, no face-to-face classes Student interacts with a teacher via her 6 virtual devices office meeting. it was Students who studied full online learning Achieved higher median values compared to blended e-learning Both Overall Continuous Assessment Score (OCAS) and Overall Inspectable Score (OES) after moderation but before that Moderate mean OCAS scores were lower when fully online Learning compared to blended e-learning. since the final Rank Score (FRS) is obtained from moderated OCAS and Adjusted OES concludes that students are learning entirely online These achievements had a higher median. A high proportion of high-risk students Blended eLearning compared to purely online learning institutions we can find them by looking at a range of online study courses for a useful result.

III. PROPOSED SYSTEM

The problem of student performance prediction is done by using machine learning algorithms and the

data is analyzed including demographic, engagement, and performance of a student. The data collection of interactions with the study materials, and how well the assignments are completed on time even if the lifestyle is not suitable but completing the work on time also gives a good performance rate. For the purpose of deeper analysis, the interaction with the virtual learning environment of students is analyzed with the student assessments taken for the module, here no previous qualification is required to complete the module in this open university and gives us to check students' performance rate.

IV. DATA COLLECTION AND PREPARATION

The dataset for this machine learning project was provided by the Learning Analytics Research Group of the Knowledge Media Institute at The Open University. This dataset is publicly available and includes student demographics, modules completed, module start times (module presentations), information about student academic performance (term papers and exam grades), and student and student Consists of tables that contain information about the interaction of University Virtual Learning Environment (VLE).

A. Assessments Info

1. Code Module: The identification code for the module to the assessment belongs.
2. Code Presentation: The identification code for the presentation to the assessment belongs.
3. ID Assessment: The identification number for the assessment.
4. Assessment Type: Tutor Marked Assessment (TMA), Computer Marked Assessment (CMA), Final Exam (Exam).
5. Date: Final submission information regarding the final submission date of the exam is calculated from the number of days from the start of the presentation of the module. The presentation start date starts from 0. (zero).
6. Weight: Rating weight (%). Exams are usually treated individually and have a weight of 100%. The sum of all other ratings is 100%. If there is no information regarding the date of the final exam, this is the end of the week of the final presentation.

Here the inconsistency of weights according to the short of the weight of exams is normally 100, and the total weight of all other assessments is 100. As a result, a module with just one exam would have a weight of 100, while a module containing one exam as well as other assessments would have a weight of 200. IDs are categorical, the Assessments IDs are denoted as integers, which is incorrect.

B. Assessments Results

1. ID Assessment: The identification number for the assessment.
2. ID Student: The unique identification number for the student.
3. Date Submitted: The student submission dates are measured as the number of days from the start of the module presentation.
4. Is Banked: The status flag indicates the review of the results which has been carried over from the previous presentation.
5. Score: The student scores in the assessment and it ranges 0-100, which is a score below 40 and it is interpreted as a failure and the grades range from 0 to 100.

Here the inconsistency is the student who does not submit the assessment, no result is recorded. Therefore, all null scores can be interpreted as non-submissions. This means we can fill them out with zeros.

C. Courses Info

1. Code Module: The code name for the module which serves as an identifier.
2. Code Presentation: The code name for the presentation. Announcements from February are composed of the year and "B", and announcements from October are composed of "J".
3. Length: The duration of the module presentation is in days.

Here no inconsistency was found.

D. Student Registration

1. Code Module: The Module identification code.
2. CodePresentation: An identification code for the presentation.

3. IDStudent: The student's unique identification number.
4. Date Registration: The date the student registered for the module presentation. This is the number of days counted from the start of the module presentation if there is a negative value which means the student registered for the module presentation 30 days before the start.
5. Date Unregistration: The module presentation unenrolment date is the number of days relative to the start of the module presentation. Students who have completed the course will have this field blank. A student who has not enrolled has the withdrawal as the value of the final result column in the student Info file.
6. Weight: Rating weight (%). Exams are usually treated individually and have a weight of 100%. The sum of all other ratings is 100%. If there is no information regarding the date of the final exam, this is the end of the week of the final presentation.

Here is the inconsistency of the date of un-registration: The student does not submit the assessment, and no result is recorded. Therefore, all null scores can be interpreted as non-submissions. This means we can fill them out with zeros.

E. VLE Resources

1. ID Site: An identification number for the material.
2. Code Module: An identification code for the module.
3. Code Presentation: An identification code for the presentation.
4. Activity Type: A role associated with the material of the module.
5. Week From: A week that the material is expected to be used.
6. Week To: The use of material that is planned to use the materials.

Here the inconsistency of the date of a week from and week to is with the student who does not submit the assessment, no result is recorded. Therefore, all null scores can be interpreted as non-submissions. This means we can fill them out with zeros

F. VLE Interactions

1. Code Module: An identification code for the module.
2. Code_Presentation: An identification code for the module.
3. ID_Student: An identification number that is unique for the student.
4. ID_Site: The VLE material identification number.
5. Date: The date that the student interacted with the material which was measured as the number of days since the presentation of the module began.
6. Sum Click: How often students will interact with the material during the day.

No inconsistency was found.

G. Student Information

1. Code_Module: The identification code for the module to the assessment belongs.
2. Code_Presentation: The identification code for the presentation to the assessment belongs.
3. ID Student: The student identification number which is unique.
4. Gender: The gender of the students.
5. Region: Identify the region where the student lived during the presentation of the module.
6. Highest Education: The highest qualification of the student when they are attending the module presentations.
7. IMD Band: Provides an index of multiple depravity bands where the students lived during the presentation of the module.
8. Age Band: Provides the student's age.
9. Number of previous attempts: The number of times that the student has attempted the module.
10. Studied Credits: The module total credits where the student is studying currently.
11. Disability: An indication reported of the disability of a student.
12. Final Result: Final student results in module presentation.

Here is the inconsistency of the average rank is a holistic measure that is allocated to an area based on both its performance (score) and how this performance compares to other areas on a national scale.

V. DATA CLEANING AND PREPROCESSING

Finding the inconsistencies of the weights The descriptions typically state that the exam weight is 100, and the sum of all other scores is 100. This means that a module with only one exam will have a weight of 100, and a module with one exam and some assessments will have a weight of 200. Summing the rating weights and grouping by the presentation of the module. After the above process clearly, we observe that the total weight of the module presentation is 200, excluding the module CCC of 300 and the module GGG of 100. Let's take a closer look.

So, we see the weights of the exam in what way they are in the module presentation. Now after the above the modules have a weight of 100 in an exam except the CCC module in each of the module presentations. So, now we count the number of exams included in each of the presentations of the module. Now, The CCC module has two exams, which explains why this module is rated so highly. So, looking at all the non-exam tasks and see if everything is correct or not. So, now we sum up all learning outcomes of the weights for each of the module presentations. Now, here we can see that the GGG module has no weight in its allocation. Checking whether it is because of the no tasks for this module. Checking whether there are other TMA, CMA, allocations with a weight of 0. We found 28 and 1 respectively. Fixing the inconsistencies of weights. The CMA assignments often have a weight of 0, so for simplicity, we assign only a total weight of 100 to TMA assignments. Checking for the Non-submissions. If a student does not submit an assessment, no score will be recorded. Therefore, any zero scores can be interpreted as non-submission. That is, you can pad them with zeros. But it's a bit strange that reviews with zero value have a due date. For unsubmitted reviews, the Submission Date value is expected to be zero. Ideally, this should be made clear to your data supplier.

Performing the below by merging for our analysis:
VLE + VLE Interactions
Registration info + Courses + Info
Assessment info + Assessment Results

Calculating the weighted score:
Late submission, Fail rate, Date registration, Total clicks, Weighted Score, Late Rate, Fail Rate.

VI. EXPLORATORY DATA ANALYSIS

A. Univariate and bivariate analysis

1. Distribution plots: The dataset has many skewed variables. Care should be taken when using linear models, as they assume a normal distribution.
2. Distribution score per module: The target variable has two peaks, is not normally distributed, but has no outliers. Ultimately, you may want to transform your target to improve your model. However, since this notebook demonstrates a very basic analysis, it does not do that, but you should be careful when using certain models (such as linear regression, which assumes the distribution is normal).
3. Correlation matrix: Indicates that there is little collinearity between the variables. Let's take a closer look at the linear correlation between features and targets.
4. Target correlation: The weighted score is most positively correlated with a total click. The more students use Blackboard, the better the results. Also, there is a weak negative correlation with the number of past trials. Late rate and fail rate also showed a weak but negative correlation with weighted scores. No correlation was found with the module presentation length or date registration, or studied credits.
5. Univariate analysis of data: Very few students with no formal education. Very few students with post-grad qualifications. All other variables vs the final score. The results are ambiguous. There doesn't seem to be a strong relationship between variables and goals, except maybe total clicks.
6. Frequency count of the different modules by a pass rate: We can observe the one which shows the frequency count of a pass rate is more with the module AAA and GGG but module CCC has more failure rate.
7. Distribution of scores per module: We can observe that some modules seem to have a higher failure rate than others. The box plot also shows some outliers.
8. Frequency count of gender by a pass rate: We can observe the one which shows the frequency count

of a pass rate has a little difference from the gender male and females has less failure rate.

9. Distribution of scores per gender: We can observe that the weighted score of gender males and females has little difference. The box plot shows no outliers.
10. Frequency count of the region by a pass rate: We can observe that the one which shows the frequency count of a pass rate is more in northwestern region and Wales region.
11. Distribution of scores per region: We can observe from the one that the weighted score of regions has very little difference. The box plot shows no outliers.
12. Frequency count of highest education by a pass rate: We can observe the one which shows the frequency count of the pass rate of students that have a lower than a level the previous education seems to fail more.
13. Frequency count of IMD band by a pass rate: We can observe the one which shows the frequency count of the pass rate of students with different deprivation bands has very little difference and 0-10 percent falls among the most deprived small areas and we can observe the increasing of the band the area of deprivation is more.
14. Frequency count of age band by a pass rate: We can observe the one which shows the frequency count of the pass rate of students has little difference where the grouping age of 0-35 and over 35 is been done.
15. Frequency count of disability by a pass rate: We can observe the one which shows the frequency count of the pass rate of students with disability has a little more failure rate than non-disable ones.

VII. METHODOLOGY

A. Data Models

1. Regression: A regression version affords a characteristic that describes the connection among one or extra unbiased variables and a response, dependent, or goal variable. A regression analysis gives the prediction and determining the effects on the target variable. Setting an encoding and scaling instructions and transforming to the features.
2. Linear Regression: Linear regression is a model in which the relationship between input and output is a straight line. This is the easiest to design and can even be observed in the real world. Even when relationships are not very linear, our brains recognize patterns and try to fit a basic linear model to them.
3. LASSO Regression: Lasso regression is a technique which is commonly used in machine learning to select a subset of variables, it provides higher predictive a to other regression models. Lasso regularization helps improve model interpretation.
4. Support Vector Regression: The support vector regression is used for predicting discrete values. Support vector regression uses the same principles as SVM. The basic idea behind SVR is finding the best line. In SVR, the best-fit straight line is the hyperplane with the maximum number of points.
5. Decision Tree: Decision trees are used to add noisy observations to approximate a sine wave. As a result, we learn a local linear regression that approximates a sine wave.
6. GradientBoost: Gradient boosting which is the Gradient Boosted Regression Trees (GBRT) which is a flexible nonparametric statistical learning technique for classification and regression.
7. K Nearest Neighbours Regression: A onparametric technique that intuitively approximates the relationship between an independent variable and a continuous outcome by averaging observations in the same neighborhood.
8. Random Forest: Random Forest Regressor is a random forest which is a meta-estimator that fits a set of classification decision trees to different subsamples of a dataset and uses averaging to improve prediction accuracy and control overfitting.

B. Data Models

1. Decision Tree: Decision trees are nonparametric supervised learning algorithms used for both classification and regression tasks. It has a hierarchical tree structure consisting of root nodes, branches, inner nodes and leaf nodes.

2. Random Forest: A classification algorithm that consists of many decision trees. In creating the individual trees, we use bagging and feature randomness to try to create a forest of uncorrelated trees in which the predictions of the committee are more accurate than the individual trees.
3. Support Vector Machine: A supervised machine learning model that uses a classification algorithm for two-group classification problems. After giving the SVM model a set of categorically labeled training data, new text can be classified.
4. Stochastic Gradient Descent: Stochastic Gradient Descent SGD Classifier which is a linear classifier which is a simple but highly efficient approach for fitting linear classifiers and regressors to convex loss functions such as (linear) support vector machines and logistic regression.

VIII. MODEL SELECTION, MODEL BUILDING AND EVALUATION

Evaluation of the above model uses randomized training and test set splits and computes RMSE and adjusted R2. The adjusted R2 shows that the model explains 35% of the total sample variance. An RMSE of 23.8 indicates an error. The forecast is off by 23.8 points. Considering the error rate is only 40%, this is a big mistake. Performing cross-validation the Cross-validation uses stratified k-fold cross-validation that differs from randomized value validation. The training set is split into a small training set and a smaller validation set. Each of these sentences is used in turn for training and validation. Cross-validation also shows that the model performs poorly, as expected.

Table A. Regression Model Performance

Models	RMSE	Adjusted R2 score	Mean	SD
Decision Tree	17.15	0.66	20.41	0.49
Gradient Boost	17.73	0.64	18.39	18.39
KNN	16.85	0.67	29.06	0.33
Random Forest	16.97	0.67	18.56	0.24

Table B. Classification Model Performance

Models	Score	SD
Random Forest	0.78	0.0029
SVM	0.77	0.0022

GradientBoost and Random Forest are the best models. Cross-validation of the GBoost model shows it to be the most accurate model, but the training set error without cross-validation is higher for GradientBoost than for RF. ANN had the lowest RMSE score on the training set without cross-validation.The support vector machine classifier model performed the best. the SVC model performs with slightly less variance between the scores during cross-validation as shown by lower standard deviation.

CONCLUSION

The modules offered by the open university with materials that can be referred to the activities of a student by the usage and the fulfillment of the requirements with numerous discrepancies where predicting which students will drop out or fail their modules and which will pass them is the current task at hand by finding the riskiest failure of a student in their module. Best Regression Model Gradient Boost and Random Forest are our best models, Even though the error for the training set without cross-validation for the Gradient Boost model is higher than for the RF, cross-validation for the Gradient Boost model demonstrates that it is the most accurate model. Without cross-validation, KNN had the lowest RMSE score for the training set, but we can see how this model's performance declined during cross-validation, indicating that it is overfitting. And the Best Classification Model The classifier model using a support vector machine performed the best. Although the RF model's accuracy scores of 0.78 and the SVC model's accuracy scores of 0.77 are close, the SVC model performs with a little less score variance during cross-validation, as indicated by a smaller standard deviation. Both the training and test sets generate nearly identical accuracy scores. The best model for the regression task was Gradient Boost which yielded RMSE = 17.73 and adjusted R2 = 0.64and mean value of 19.39 when evaluated on the test set. The best model for the classification task

was the support vector machine classifier 0.77 accuracy score on the test set.

REFERENCES

[1] Kuzilek J., Hlosta M., Zdrahal Z. Open University Learning Analytics dataset Sci. Data 4:170171 doi: 10.1038/sdata.2017.171.

[2] S. Voghoei, N. Hashemi Tonekaboni, D. Yazdansepas and H. R. Arabnia, "University Online Courses: Correlation between Students' Participation Rate and Academic Performance," 2019 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2019, pp. 772-777, doi: 10.1109/CSCI49370.2019.00147.

[3] Y. Widyaningsih, N. Fitriani and D. Sarwinda, "A Semi-Supervised Learning Approach for Predicting Student's Performance: First-Year Students Case Study," 2019 12th International Conference on Information & Communication Technology and System (ICTS), Surabaya, Indonesia, 2019, pp. 291-295, doi: 10.1109/ICTS.2019.8850950.

[4] Liz-Domínguez, Martín, Manuel Caeiro-Rodríguez, Martín Llamas-Nistal, and Fernando A. Mikic-Fonte. 2019. "Systematic Literature Review of Predictive Analysis Tools in Higher Education" Applied Sciences 9, no. 24: 5569. https://doi.org/10.3390/app9245569

[5] R. C. Raga and J. D. Raga, "Early Prediction of Student Performance in Blended Learning Courses Using Deep Neural Networks," 2019 International Symposium on Educational Technology (ISET), Hradec Kralove, Czech Republic, 2019, pp. 39-43, doi: 10.1109/ISET.2019.00018.

[6] Alshareef, Fatima, et al. "Educational data mining applications and techniques." International Journal of Advanced Computer Science and Applications 11.4 (2020).

[7] Križanić, Snježana. "Educational data mining using cluster analysis and decision tree technique: A case study." International Journal of Engineering Business Management 12 (2020): 1847979020908675.

[8] S. Batool, J. Rashid, M. W. Nisar, J. Kim, T. Mahmood and A. Hussain, "A Random Forest Students' Performance Prediction (RFSPP) Model Based on Students' Demographic Features," 2021 Mohammad Ali Jinnah University International Conference on Computing (MAJICC), Karachi, Pakistan, 2021, pp. 1-4, doi: 10.1109/MAJICC53071.2021.9526239.

[9] E. Buraimoh, R. Ajoodha and K. Padayachee, "Application of Machine Learning Techniques to the Prediction of Student Success," 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), Toronto, ON, Canada, 2021, pp. 1-6, doi: 10.1109/IEMTRONICS52119.2021.9422545.

[10] F. Alnassar, T. Blackwell, E. Homayounvala and M. Yee-king, "How Well a Student Performed? A Machine Learning Approach to Classify Students' Performance on Virtual Learning Environment," 2021 2nd International Conference on Intelligent Engineering and Management (ICIEM), London, United Kingdom, 2021, pp. 1-6, doi: 10.1109/ICIEM51511.2021.9445286.

[11] S. K. Yeung and W. L. Lee, "Full Online Learning and Blended e-Learning: A Comparison of Students' Performance," 2019 IEEE International Conference on Engineering, Technology and Education (TALE), Yogyakarta, Indonesia, 2019, pp. 1- 7, doi: 10.1109/TALE48000.2019.9225871.

[12] Qiu, F., Zhang, G., Sheng, X. et al. Predicting students' performance in e-learning using learning process and behaviour data. Sci Rep 12, 453 (2022). https://doi.org/10.1038/s41598-021-03867-8

[13] Bilal, M., Omar, M., Anwar, W. et al. The role of demographic and academic features in a student performance prediction. Sci Rep 12, 12508 (2022). https://doi.org/10.1038/s41598-022-15880-6