Recent Advancements in Prediction of Early Dropouts from Online Courses using Self-Regulated Learning

Harika Dondapati[1], Lakshmi Priya Komalli[1], Tejaswi Koneru[1], Jaynica Nunna[1], Sumalatha Saleti[1]

Department of Computer Science and Engineering, SRM University – AP, Andhra Pradesh, India

Email: harika_dondapati@srmap.edu.in

lakshmipriya_komalli@srmap.edu.in

tejaswi_koteswararao@srmap.edu.in

jaynica_nunna@srmap.edu.in

Abstract: Massive Open Online Courses (MOOCs) typically have significant dropout rates. To address this problem, many studies have suggested algorithms for early detection of learners. However, they don't take expertise in various factors, like Self-Regulated Learning (SRL), which can significantly impact the success of students. Furthermore, predictions are frequently made in instructor-paced content released over time in MOOCs, but not in self-paced MOOCs, where users are free to enrol, and all resources are accessible right away. This research contributes to resolving these problems by exploring the integration of SRL methods into self-paced MOOC predictive models. Self-reported and event based SRL techniques are assessed and contrasted to determine how well they predict dropout rates. An efficient dropout prediction model can pinpoint the contributing causes and offer guidance on how to launch interventions to boost students' success in MOOC.

Keywords: Machine learning Models; Online Courses; Instructor-paced; Self-paced; MOOC; Self-Learning; Self-Evaluation; Self-Monitoring;

INTRODUCTION

Massive Open Online Courses (MOOCs) are the courses that institutions provide on specific MOOC sites (e.g., edX and Coursera). Anyone with access to the Internet can take these courses for no cost. Usually, no prior knowledge is necessary. MOOCs frequently draw hundreds of students to thousands because of their accessible nature. MOOCs [Daniel, 2012; Sahin, 2021] gave an opportunity to all global learners to avail quality education through online mode. But considering these days there is a high dropout rate problem. This made a serious issue on education too. Many people who join MOOCs are confused about

whether to complete their course or not. Therefore, predicting the dropout in MOOC is the main topic to research in recent years [Adnan et al., 2021; Alarpe et al., 2022; Jansen et al., 2020; Renzhe 2021]. In online courses such as MOOC, technical support for self-regulated learning is very rare. Self-evaluation is one of the main things that one can know about learning through online courses. Although studies have shown that students' use of self-regulated learning varies by learning environment, such as online versus physical settings, they were once thought to be somewhat stable across learning situations [RahulKatarya et al., 2021]. Self-Regulated Learning (SRL) [Broadbent & Poon, 2015; Jansen et al., 2020; John et al., 2022; Kizilcec et al., 2016] is based on evidence of actions taken during learning and changing the learning procedure is carried out through online courses. Nowadays, MOOC learners are increasing globally but the completion rates of chosen courses are very low and dropout rates are very high**.** According to many case studies, the completion of MOOCs is maximum from 5% to 10%, whereas the dropout rates are high at 80 to 90% [Goldberg et al., 2015]. Many students who are dropping out from the MOOCs or any online course are because they are not fully connected through online mode. The quality of education that is provided online is not accurate such as assignments, implementation of the courses, and online support is not possible all the time [Bhutto et al., 2012]. There is a lack of understanding of learning behavior through the online mode of education. The results after the course completion are very low because there is a lack of interaction between the learners and the instructors. The students who are pursuing their courses online are in large numbers, so the information is stored in large data, including the student ID, years, enrollment date and time, course ID, gender, the number of courses, and many more features. For students to succeed in online learning, particularly in settings with little support and direction like MOOCs, Self-Regulated Learning is crucial. It is a set of abilities that can be learned and enhanced through experience and repetition.

When the learning process is less externally regulated, SRL becomes more crucial for learner success. SRL is split into five phases: 1. Process 2. Purposeful 3. Forethought 4. Self-Monitoring 5. Self-Reflection. In the Process phase, self-regulating learners organize their learning and create goals. In the second learning phase - Purposeful planning, the author needs to plan steps that point to the end, and the third phase is forethought, self-regulating students develop objectives and organize their studies. The Self-Monitoring phase is like watching what they're learning i.e., how am I doing today and getting assistance when necessary and concentrating their efforts on choosing the right tasks and how to approach them in order to eventually accomplish a given objective. For instance, students must establish the sequence of events, their timing, and procedures for accomplishing goals, such as the method and effort. In phase four, attention control notes to pay self-attention towards the end. The final phase which is self-reflection is about evaluating the learning process and its progress. As a result, they can finally

conclude regarding their learning process and enhance their strategies for applied learning. Self-learning learners are more engaged, and volunteer answers perform better on tasks that are more in conflict. Numerous studies have been conducted on the impact of SRL on academic performance and course results. The associations between SRL and academic success, and SRL and course outcomes have been established through conceptual to be considerable and favorable across educational levels.

RELATED WORKS

Since SRL encompasses cognitive, metacognitive, motivational, behavioral, and affective processes, it serves as a catch-all word for a variety of factors that are examined. Numerous models have been put forth in the literature to summarize all the factors involved in the SRL process due to its complexity [De Barba et al., 2016]. Despite their various conceptualizations, most researchers view SRL as a cyclic three-phases of the process: preparation, execution, and evaluation. The preliminary phase is preparation that is extremely important, and incorporates SRL techniques like goal setting, strategy planning, goal orientation, task motivation, and other factors influenced by the learner's beliefs and motivation value, anticipated results, and self-efficacy. Motivating people is directly tied to SRL. According to De Barba et al., (2016), motivation was a significant predictor of performance in MOOCs and a significant factor in dropout rates. In MOOC environments, motivation and SRL skills typically go hand in hand since motivated students typically report having greater SRL skills. However, the MOOC context can affect motivation and even change it. Keeping students motivated throughout the course could lower dropout rates among those who really want to finish the MOOC. The performance phase i.e., execution, where both cognitive and metacognitive processes occur, includes SRL approaches including time management, study environment management, and aid seeking. The assessment phase i.e., evaluation, where metacognitive processes predominate, includes SRL approaches like self-reflection, self-evaluation, or self-satisfaction.

It's interesting to note that Kizilcec et al. (2016) claimed that learners could be motivated to engage in SRL. For instance, learners' SRL abilities will advance if they are motivated and at ease in the learning setting. In the context of MOOCs, it examines the student traits or aspects that may make it easier to distinguish between disparities in SRL skills and motivation levels. Instructors may be able to identify groups that are in danger of dropping out and give them the right support by using the study's insights. The research used online students from six different MOOCs to make up the final study sample.

Most of the students who enroll in online courses do not complete them and drop out midway. MOOC development might be limited by various factors. It's important to be able to predict whether a student will drop out. The data needed for the dropout prediction challenge is raw activity data of students on an online course program during the time. The challenge we must evaluate is if these students will drop out of their classes in the future. This is a topic of binary classification. Most of these solutions relate to the standard method of classification problems. The raw activity data are used to extract features in the first stage. The classification process is completed in the second stage using classification algorithms. Even though these methods have produced strong results, they have some faults. One significant issue is how characteristics are extracted in the first step. Feature extraction is commonly achieved using feature engineering in these approaches. When using feature engineering to extract features, various iterations of feature extraction and testing are required. As a result, the process takes a long time and is unreliable. The technique of manually extracting features from raw data is known as feature engineering. This procedure is done iteratively, with all features retrieved by hand. In feature engineering, on the other hand, methodologies for extracting features are adapted to the features of datasets. Strategies that work for one type of dataset may not work for another type of dataset. New algorithms for extracting features must be created if new types of datasets exist.

The authors in [Dass et al., 2021] focused on utilizing Random Forest technique to predict student dropouts. Machine Learning (ML) is a powerful technology that may be used in Learning Analytics to uncover hidden patterns of student involvement in MOOCs [AL-Shabandar et al., 2017; Kloft et al., 2014; Sarwat et al., 2022], offering advantages above standard statistical analysis. K-Nearest Neighbors (KNN), Decision Trees (DT), and Random Forest are examples of machine learning algorithms that have been used to predict students' learning success at the end of a course. The authors [Dass et al., 2021] tested a varied data store, involving official and informal educational aspects, using both statistical and predictive models.

PROPOSED SCHEME

Dataset Description

Prediction of early dropouts from online courses conducted by the student course dropout prediction dataset are included in the collection of data like student_id, years, entrance_test_score, enroll_date_time, course_id, course_progress, online, gender, international, complete. From 2016 to 2020, data from the personality math class College Algebra and Solving Problems presented by EdPlus on the MOOC program Open edX at Arizona State University (ASU) was considered. In addition to the existing features, survey information has been included. The survey consists of positive and negative questions on

SRL. The student SRL profiles can be understood using this data as a foundation for learning success prediction. As shown in Table 1, the dataset includes SRL data on the four elements of Planning, Self-monitoring, Attention Control, and Self-Evaluation.

The student demographic information was studied to acquire a sense of the students' backgrounds, and this description aids us in evaluating the research's impact. Table 2 shows the course completion / dropout percentage of students.

Table1. Factors of Self-regulated learning

S.No.	Factor	Positive Statement	Negative Statement
1	Planning	If an important test is coming up, I created a study plan	I struggle to make strategies that will help me achieve my objectives.
2	Attention Control	–	When I fall behind on my work, I frequently give up.
3	Self-Monitoring	I monitor my results in achieving my objectives	I have difficulties remembering everything I have to do.
4	Self-Evaluation	I make an attempt to learn from my experiences when I lose at something	–

Table 2. The course completion / dropout percentage of students

Course	Number of students	Percentage
Dropout	1721	12.5%
Complete	4136	87.5%

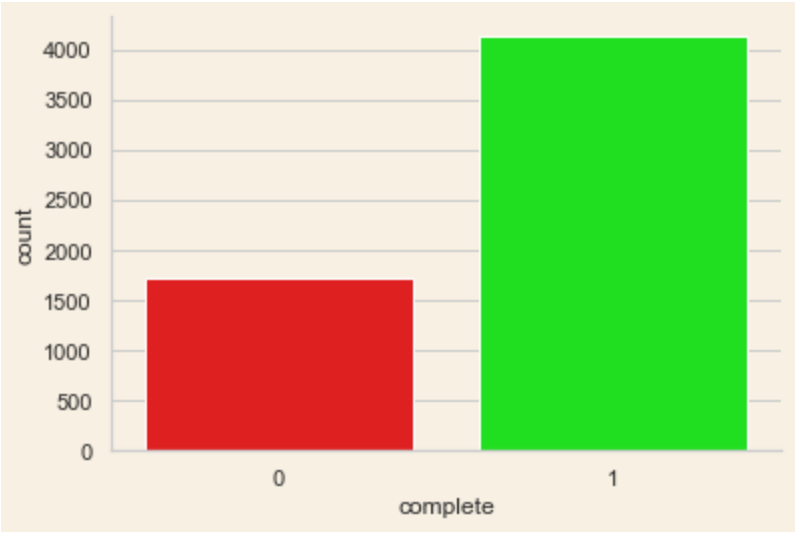

Figure 1. Distribution of Students in Courses Completion

Figure 1 depicts the course completion status of students from ASU dataset. It is observed that the percentage of course completion is high when compared to dropouts from the courses.

Data Extraction

Data extraction is the process of obtaining information from a data source using data mining techniques. The ALEKS platform has an API for accessing the data in the SQL Database. The data selection technique was used to choose a query from the data source using the data selection method. The query table is saved as a database after the data selection is completed. The python programming language is used to process the data. This course has a total of 5857 students, all have some degree of proficiency. Following their Initial Knowledge Check (IKC), they will engage in an activity. The IKC is a proficiency test that is administered to students. All students will be asked to assess their present knowledge at the start of the course. In addition, the ALEKS system adapts the student's theoretical understanding depending on the IKC and moves them forward from their current knowledge area.

Data Preprocessing

Data preprocessing is the procedure for processing original data to be used in a machine learning model. It is a critical step in cleaning and preparing data for a machine learning model, which enhances the model's accuracy and efficiency. The panda's library includes data cleaning and analysis tools. It provides features for data exploration, cleansing, transformation, and visualization. Numpy is a python programming package that allows you to handle massive multi-dimensional arrays and matrices using a vast number of high-level mathematical operations. Matplotlib is a data visualization package that allows you to quickly create frequency distribution, plots, error charts, scatter plots, and bar graphs using only a

few lines of code. Scikit-learn is a Python library that covers classification, regression, clustering, and dimensionality reduction.

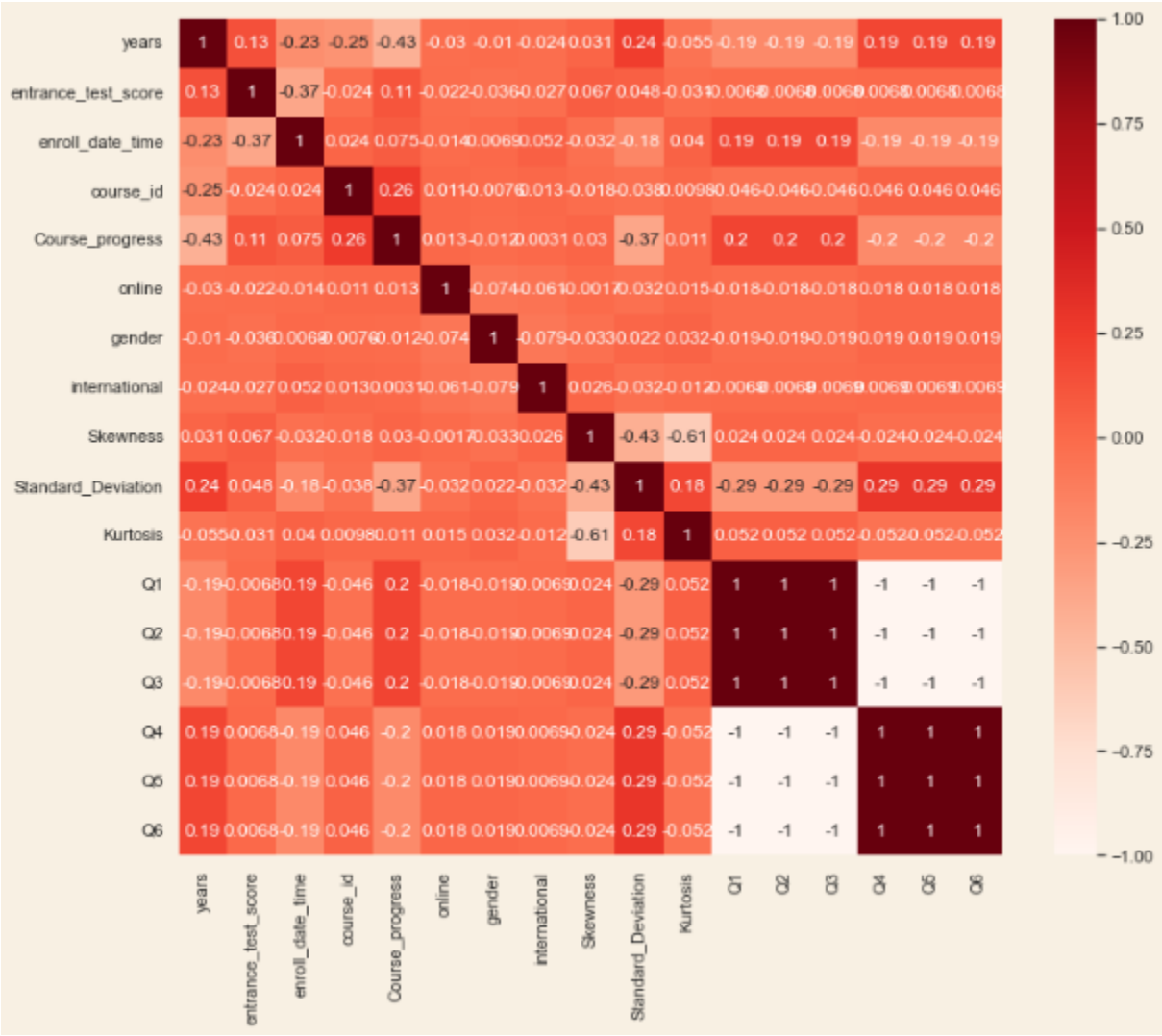

Figure 2. Correlation matrix

The correlation matrix is used for a relation with numerical columns as its input and computes the Pearson Correlation Coefficient for each pair of those columns. In addition to the existing features present in the ASU Dataset, six questionnaires have been included in the current chapter to know the effect of SRL on MOOC. The correlation matrix for the current dataset is shown in Figure 2.

The Questionnaires are given below

Q1. When I fall behind on my work, I frequently give up (N).

Q2. I have difficulties remembering everything I have to do (N),

Q3. I struggle to make strategies that will help me achieve my objectives. (N)

Q4. If an important test is coming up, I created a study plan(P).

Q5. I monitor my results in achieving my objectives(P).

Q6. I make an attempt to learn from my experiences when I lose at something(P).

Modeling and Machine Learning

Machine Learning modeling predicts MOOC dropout based on the supplied input characteristics. There are two phases to Machine Learning modeling: Model fitting and feature selection are the first steps, followed by model prediction and training**.** ML algorithms, such as Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Trees (DT), Random Forest, Gradient Boost, and ADABoost, have been applied to the prediction of students learning performance at the end of the course. Seven machine learning methods were used to examine students with a high risk of failing in order to detect their performance early. The Random Forest, Gradient Boost, Decision Trees, and AdaBoost Classifier algorithms were the most successful at identifying learners immediately. The accuracy was determined to be 96%. Furthermore, data cleaning has been proven to be helpful in improving the performance of machine learning algorithms.

Model Fitting and Feature Selection

The produced characteristics were analyzed and approved in this stage. The correlation matrix approach of Exploratory Data Analysis (EDA) was utilized to assess the characteristics in order to predict student learning outcomes in MOOCs. As a result, data leakage in the ML model would be avoided, and the modeling project's success would be increased.

Model Prediction for Testing and Training

Machine learning models are trained with the training data and are tested. The model forecasts if a student will fail the course or complete it when the model is fed information on a student's learning in a MOOC. We used trials to assess the model's performance and check if it could predict the proper outcome for a given collection of characteristics after it was trained. We conducted five trials to evaluate the model's performance in this study. Given several sets of input values, the sci-kit learn program was used to partition these vectors into training features, training labels, testing features and testing labels. 75%of the produced data was utilized to train the model, while the remaining 25% was used to test the model.

EXPERIMENTAL STUDY AND RESULT ANALYSIS

Course dropout (0) and Course completed (1) status of student’s\ performances applying different methods under different metrics shows the experimental outcomes of our model. To succeed in massive

open online courses, students must self-regulate their learning skills. Students who choose to continue their education in massive open online courses will be evaluated first based on their self-evaluation capabilities. In getting the results, one survey has been conducted from the students who completed and are about to complete these courses. First, the authors choose to acquire the students dataset and represent each student as a feature of data when using the recommended ML techniques to anticipate MOOC students prediction conditions. The students who completed the courses gave positive feedback to the questions asked, and the students who were not able to complete the course gave negative feedback to the survey. If the student completes some courses and some are pending, then there are very high chances to give feedback negatively in the survey conducted. Table 3 shows the machine learning models used in the current research with different rates of accuracy, precision, F1-score, recall, and ROC curve area.

Table 3. Performance metrics of machine learning models on ASU dataset

Model	**Accuracy (%)**	**Precision (%)**	**F1-Score (%)**	**Recall (%)**	**ROC Curve Area**	
					Test	**Train**
Logistic Regression	70.3	50	58	70	62	62
Decision Tree	96.5	97	98	100	98	98
Random Forest	96.3	97	96	96	97	100
KNN Classifier	72.7	71	82	86	76	84
SVM	70.3	50	58	70	71	71
Gradient Boost	96.5	71	71	73	98	98
ADA Boost	96	96	96	96	97	99

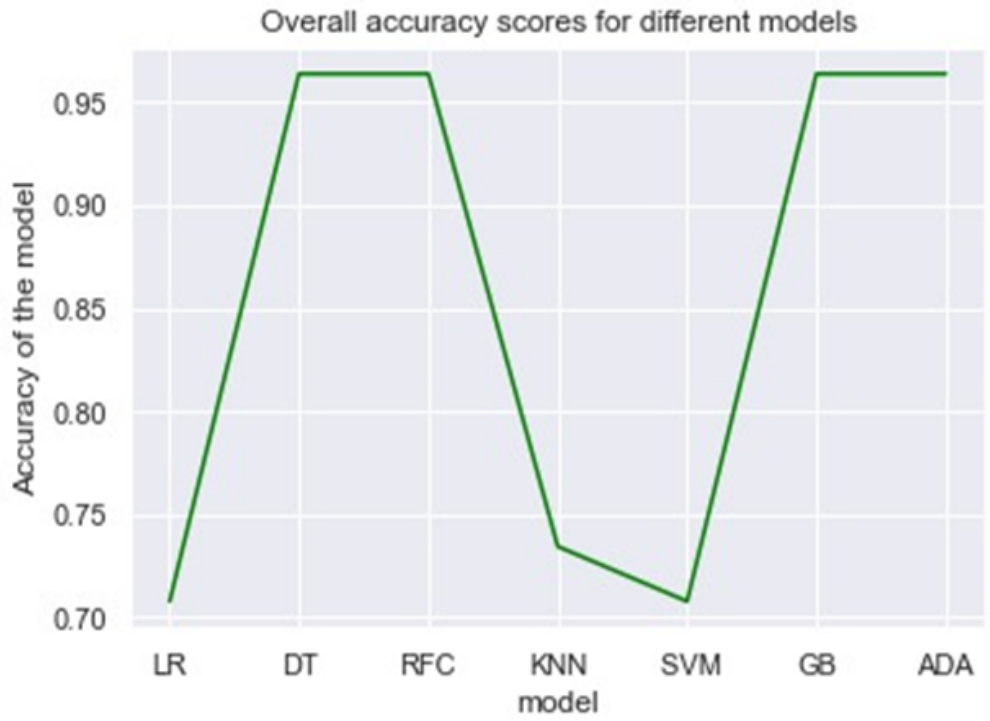

Figure 3. Accuracy Score for Different ML Models

Figure 3 describes the overall performance of accuracy scores of different machine learning models. It can be seen that Decision tree, Random Forest, Gradient Boost, and ADABoost exhibited a high accuracy of 96%.

Logistic regression

In an online learning course, logistic regression is also used to predict student dropout. Precision, recall, and accuracy were all validated using this approach. To accomplish dropout prediction based on series classification, logistic regression (LR), statistical analysis, and other methods were used depending on whether the categorization is binary or a temporal series. When it comes to classification problems, the classifier has a role in predicting dropout. To choose aspects that correlate to the most information gained, decision trees are used. After that, they employ a logistic regression model to separate dropouts from persisters. The ROC curve and confusion matrix obtained for logistic regression is shown in Figure 4. The roc curve area for train and test data is observed to be 0.62. The confusion matrix shows that 1031 records are true positive and 434 are false positive.

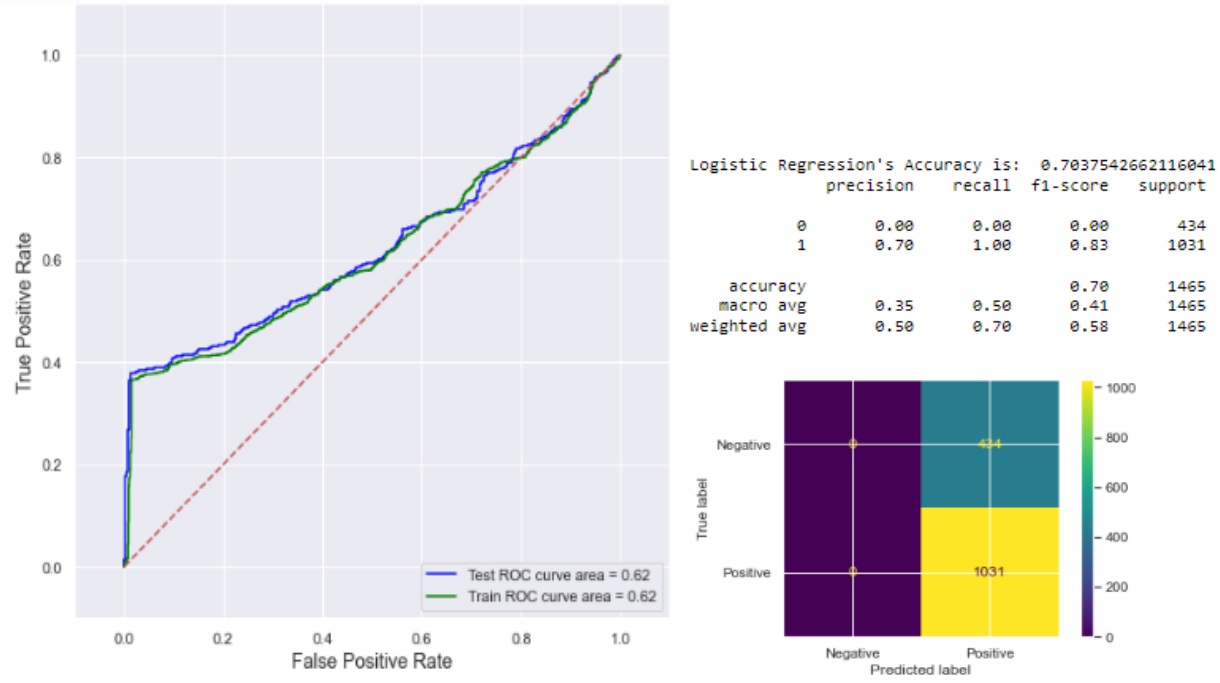

Figure 4. Roc Curve and Confusion Matrix for Logistic Regression Model

Decision Tree

Major features that benefit MOOC learners and programmers in producing course material, course design, and delivery were established using a Decision Tree (DT) algorithm. A decision tree is a map (tree-like structure) that shows the various outcomes of a decision. The Decision Tree follows a divide-and-conquer strategy. A cluster of the Decision Tree may be recognized in the Random Forest classification technique**.** The Decision Tree method was chosen as the most suited because of its extreme accuracy. The Roc curve and confusion matrix obtained for Decision Tree is shown in Figure 5. The roc curve area for train and test data is observed to be 0.98.

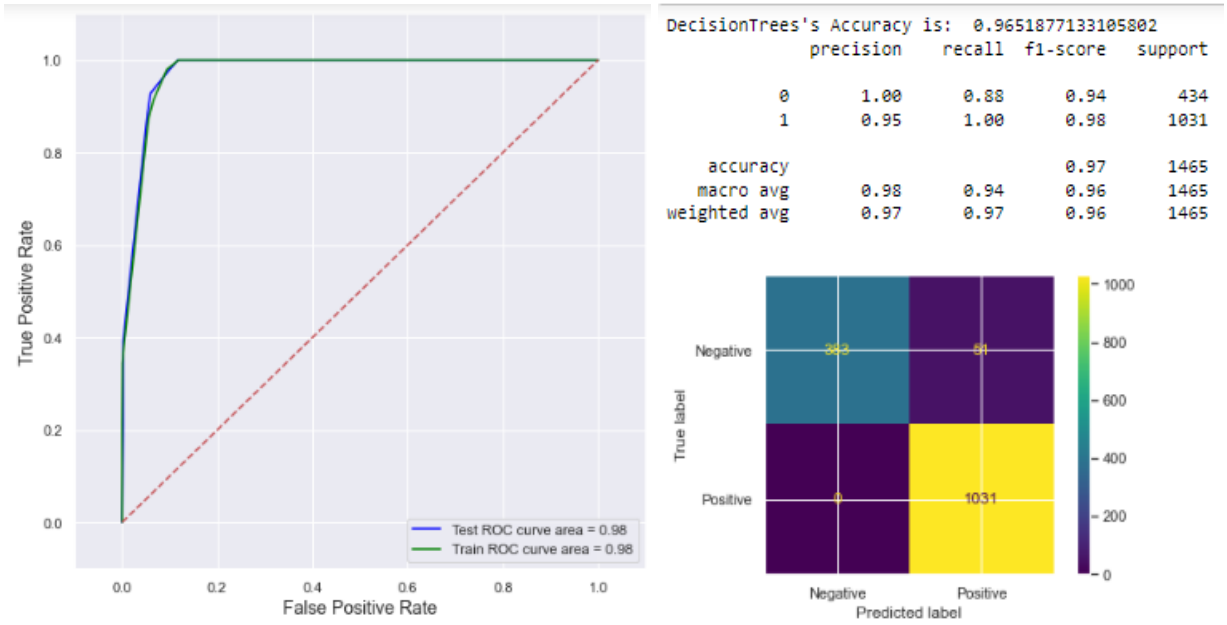

Figure 5. Roc Curve and Confusion Matrix for Decision tree Model

Random forest:

Random forest is a supervised machine learning algorithm that is commonly used to solve classification and regression applications. It creates decision trees from several samples using the majority vote for classification and the median for regression. Each tree is given a vote in order to simulate categorization. All trees are grown to their fullest potential. The forest is picked as the tree with the most votes. Preprocessing was conducted in a relevant way, and classification techniques (Decision Tree, Random Forest) were used. The RF model's accuracy is on a level with, if not greater than, that of most ML models. Outliers and noise are less noticeable in RF. Impurity is used by the classification trees in RF to decrease decision tree prediction errors. This value is reduced by RF for each tree, decreasing overfitting and data bias mistakes. As a result, RF is particularly reliable when predicting a noisy dataset with a lot of outliers. The Roc curve and confusion matrix obtained for Random forest is shown in Figure 6, the ROC curve area for train and test data is observed to be 0.97.

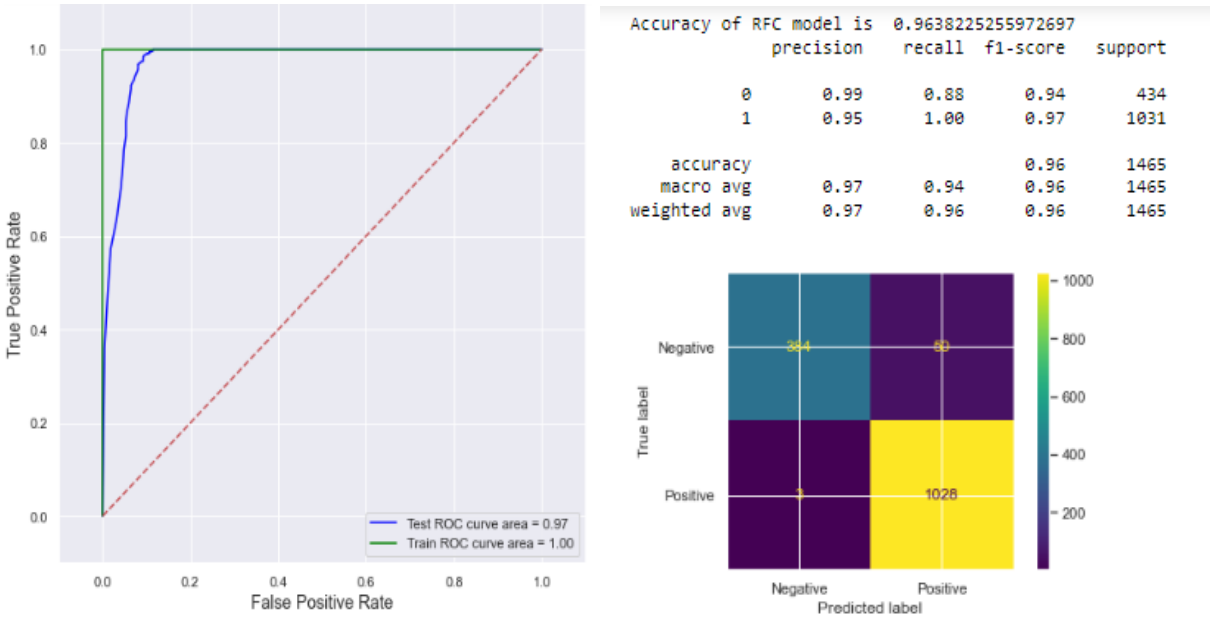

Figure 6. Roc Curve and Confusion Matrix for Random Forest Model

KNN Classifier & SVM

The K Nearest Neighbor (KNN) method is a practical and efficient classification system. For sample and class mapping, it employs the idea of a remote function. In the case of application objects with many labels, KNN is the ideal algorithm to use. It has a level of precision and consistency. SVM Dropouts can be isolated more effectively by using a weekly SVM to train. They claim that aspects from prior "history" are beneficial till a certain point in time. For each of their two models (i.e., particular and general), train an SVM classifier based on the RBF (radial basis function) kernel. and case models in general. The ROC curve and confusion matrix obtained for KNN Classifier is shown in Figure 7. The ROC curve area for train and test data is observed to be 0.76. The Roc curve and confusion matrix obtained for logistic regression is shown in Figure 8. The ROC curve area for train and test data is observed to be 0.62.

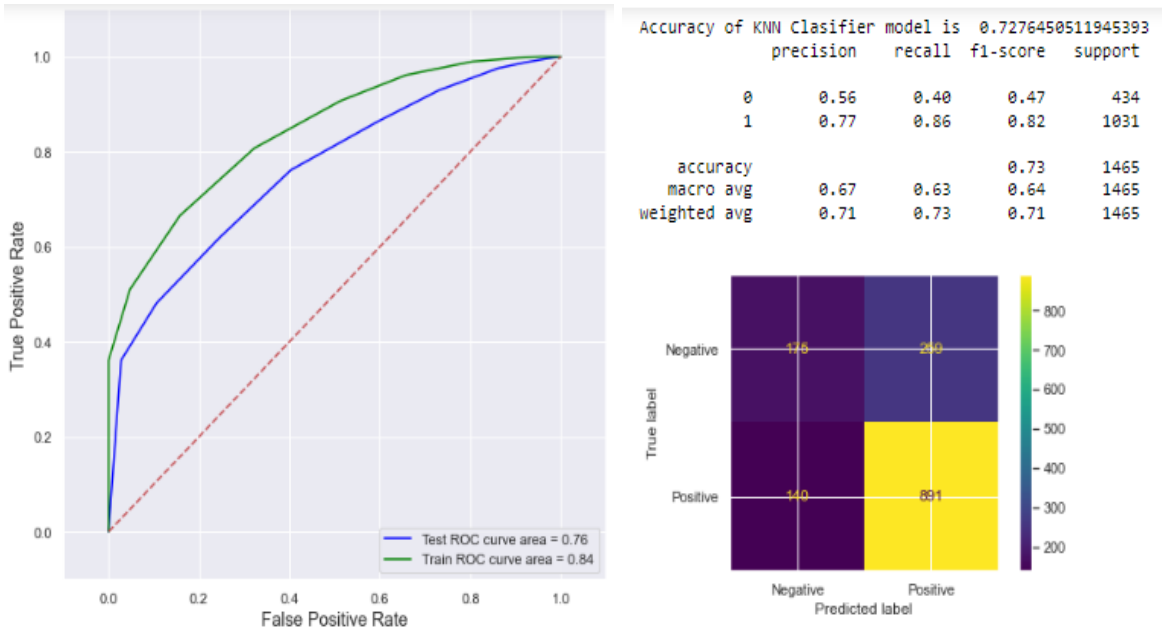

Figure 7. Roc Curve and Confusion Matrix for K-Nearest Neighbor

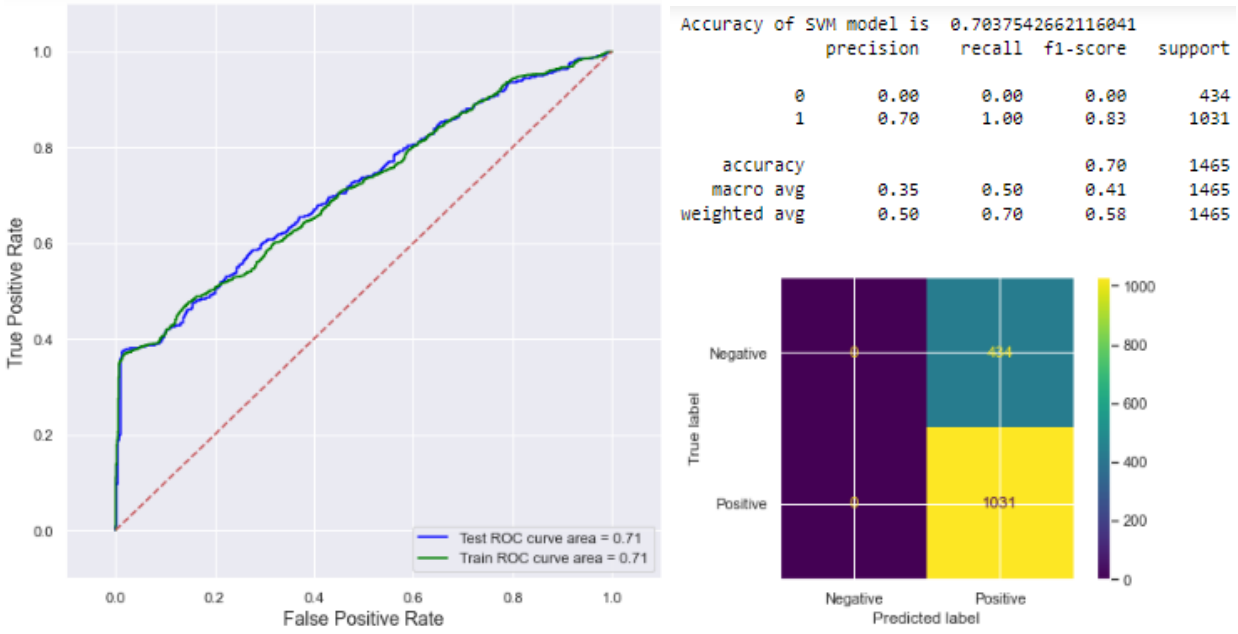

Figure 8. Roc Curve and Confusion Matrix for SVM Model

Gradient Boost & ADA Boost

To generate an overall classification model, ensemble techniques employ base classifiers (e.g., AdaBoost) and make predictions using a specific consensus function (e.g., majority voting). Random forests and AdaBoost are examples of ensemble algorithms in the SDP literature. To find the top-performing variation for their dataset, we used the Tri-Training method. Easy-to-interpret approaches such as decision

trees are pitted against non interpretable strategies. (e.g., random forests and gradient-boosted trees). The ROC curve and confusion matrix obtained for Gradient Boost is shown in Figure 9. The ROC curve area for train and test data is observed to be 0.98. The Roc curve and confusion matrix obtained for AdaBoost is shown in Figure 10. The ROC curve area for train and test data is observed to be 0.97.

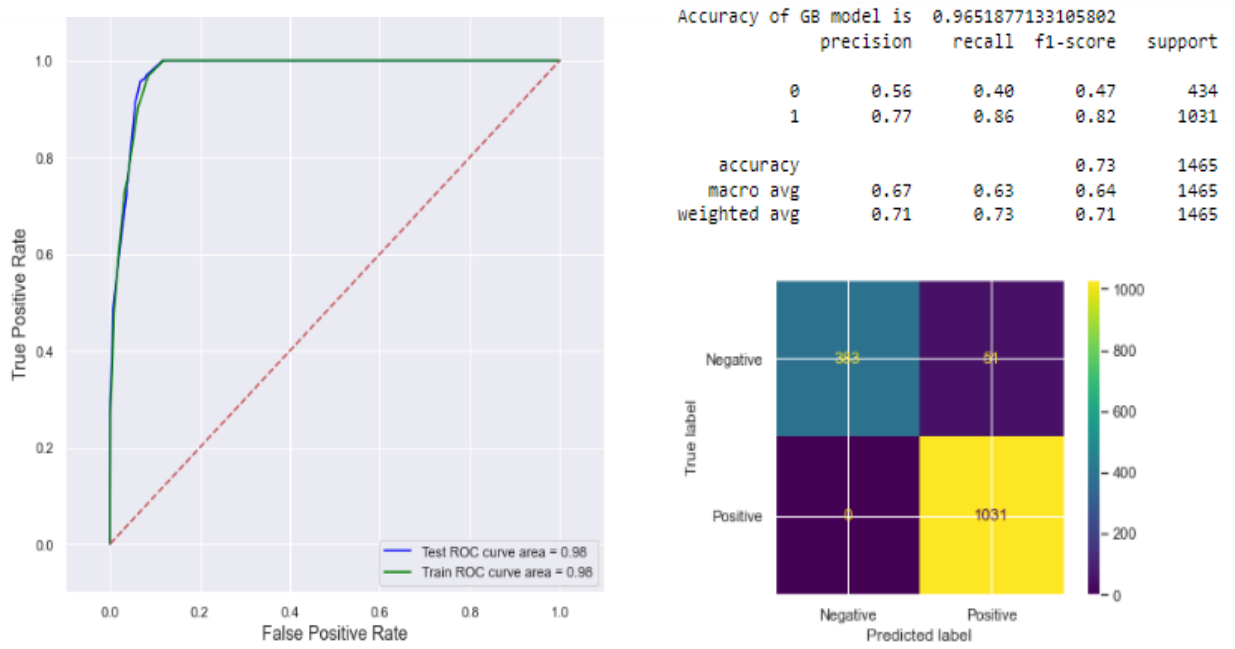

Figure 9. Roc Curve and Confusion Matrix for Gradient Boost Model

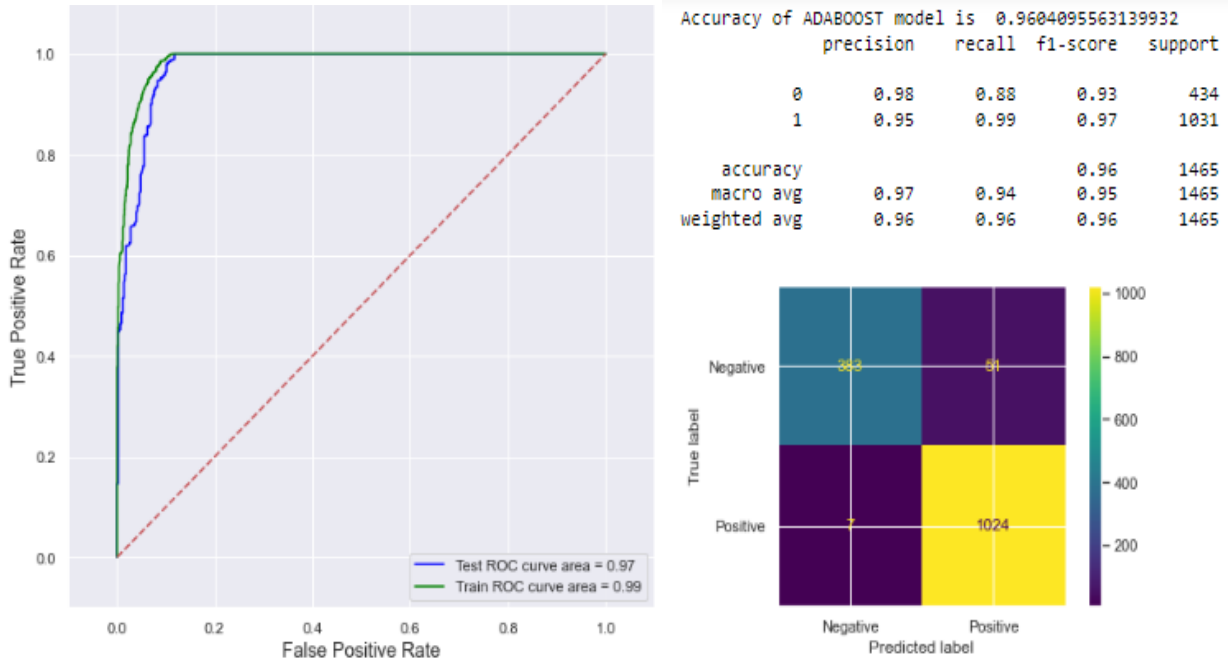

Figure 10. Roc Curve and Confusion Matrix for AdaBoost Model

CONCLUSION

The problem of predicting whether students would drop out of courses is tough for academic institutions. Furthermore, there has been little research into applying machine learning models and statistical tools to estimate retention rates in higher education. We have made a first step in detecting at-risk students early and correctly, which will help academics in developing interventions. We used several prediction models and found that regularized decision trees, Random Forest, and Gradient Boost performed best in terms of accuracy. Even though student enrolment in MOOCs has been continuously expanding, low completion rates remain a serious issue. The prediction of learner dropout will assist educational managers in evaluating and comprehending learners' learning activities based on their various interactions. It will also allow faculty members to design strategies for promoting and delivering student improvement. The findings of this study show that utilizing an ML technique can give an accurate prediction. The dataset for this study comes from Arizona State University's self-paced math course college Algebra and Problem Solving, which is available on the MOOC platform Open edX. (ASU). Accuracy, Precision, Recall, F1-score, and ROC curves are used to evaluate features and modeling in the dataset.

REFERENCES

1. Daniel, John. (2012). Making Sense of MOOCs: Musings in a Maze of Myth, Paradox and Possibility. *Journal of Interactive Media in Education, 3*.
2. Jansen, R. S., Leeuwen, A. V., Janssen, J., Conijn, R., & Kester, L. (2020). Supporting learners' self-regulated learning in Massive Open Online Courses. *Computers & Education , 146*, 103771.
3. Katarya, Rahul., Gaba, Jalai., Garg, Aryan., & Verma, Varsha. (2021). A review on machine learning based student's academic performance prediction systems, In *Proceedings of International Conference on Artificial Intelligence and Smart Systems* (pp.254-259). Coimbatore, IEEE.
4. Kizilcec, René F., Pérez-Sanagustín, Mar., & Maldonado, Jorge J. (2016). Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses. *Computers & Education*, *104*, 18-33.
5. Lynette R Goldberg., Erica Bell., Carolyn King., Ciaran O'Mara., Fran McInerney., Andrew Robinson., & James Vickers (2015). Relationship between participants' level of education and engagement in their completion of the Understanding Dementia Massive Open Online Course. *BMC Medical Education, 15*, 60.
6. Bhutto, Engr., Siddiqui, Isma., Ali, Qasim., & Anwar, Maleeha. (2020). Predicting students' academic performance through supervised machine learning. In *Proceedings of the 2020*

International Conference on Information Science and Communication Technology. Karachi, IEEE.

7. De Barba,P.G., Kennedy, G.E., & Anley, M.D. (2016). The role of students' motivation and participation in predicting performance in a MOOC. *Journal of Computer Assisted Learning*.
8. Dass, S.; Gary, K.; & Cunningham, J. (2021). Predicting Student Dropout in Self-Paced MOOC Course Using Random Forest Model. *Information*, *12*, 476.\
9. AL-Shabandar, Raghad., Hussain, Abir., Laws, Andy., Keight, Robert., Lunn, Jan & Radi, Naeem. (2017). Machine learning approaches to predict learning outcomes in Massive open online courses. In *proceedings of the International Joint Conference on Neural Networks* (pp. 713-720). Anchorage, IEEE.
10. Kloft, M., Stiehler, F., Zheng, Z., & Pinkwart, N. (2014). Predicting MOOC dropout over weeks using machine learning methods. In *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs* (pp. 60-65). Doha, Association for Computational Linguistics.
11. Cameron C. Gray., & Dave Perkins. (2019). Utilizing early engagement and machine learning to predict student outcomes. *Computers & Education, 132,* 22-32.
12. Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A.A., Abid, M., Bashir, M., & Khan, S.U. (2021). Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models. *IEEE Access, 9,* 7519-7530.
13. Sarwat, S., Ullah, N., Sadiq, S., Saleem, R., Umer, M., Eshmawi, A. A., Mohamed, A., & Ashraf, I. (2022). Predicting Students' Academic Performance with Conditional Generative Adversarial Network and Deep SVM. *Sensors, 22* (13).
14. Renzhe, Yu., Hansol, Lee., & Kizilcec, René F. (2021). Should College Dropout Prediction Models Include Protected Attributes? In *Proceedings of the Eighth ACM Conference on Learning* (pp. 91-100). New York, Association for Computing Machinery.
15. John, Saint., Yizhou, Fan., Dragan, Gašević., & Abelardo Pardo. (2022). Temporally-focused analytics of self-regulated learning: A systematic review of literature. *Computers and Education: Artificial Intelligence, 3*, 100060.
16. Şahin, M. A. (2021). Comparative Analysis of Dropout Prediction in Massive Open Online Courses. *Arabian Journal of Science and Engineering, 46*, 1845–1861.
17. Alarape, M.A., Ameen, A.O., & Adewole, K.S. (2022). Hybrid Students' Academic Performance and Dropout Prediction Models Using Recursive Feature Elimination Technique. In: Saeed, F., Al-Hadhrami, T., Mohammed, E., Al-Sarem, M. (Eds*) Advances on Smart and Soft Computing. Advances in Intelligent Systems and Computing* (pp. 93-106). Singapore, Springer.

18. Broadbent, J., & Poon, W.L. (2015). Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *The Internet and Higher Education, 27*, 1-13.