Prediction of early dropouts from online courses

Submitted in partial fulfillment of the requirements
for the award of the degree of

BACHELOR OF

TECHNOLOGY

Submitted by

Harika Dondapati AP19110010499

Research Supervisor

Dr. Sumalatha Selati

College emblem

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SRM UNIVERSITY AP, ANDHRA PRADESH

GUNTUR– 522502, INDIA

APRIL 2022

Abstract: In massive open online courses, predicting student performance is essential in order to benefit from many phases such as learning outcomes and make timely improvements. We proposed Decision trees, Random forest, Knn, Polynomial SVM, AdaBoost, Gradient Boost, and Logistic Regression algorithms, with the use of these Machine Learning Models. We have developed a project on the prediction of early dropouts from online courses, To automatically identify features from MOOCs' actual data and estimate whether each student may drop out or finish the course. We highlight some solutions that have been used to address the dropout problem, provide an analysis of the challenges of prediction models and offer some valuable insights and recommendations that could lead to the development of effective and useful machine learning solutions to address the MOOC dropout problem.

Index Terms
Machine learning Models, Online Courses, Dropouts, Prediction.

Introduction:

Massive Open Online Courses(MOOCs) have given 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 joined the MOOCs are confused about whether to complete their education or not. Therefore, predicting the dropout in MOOCs is the main topic to research in recent years. Nowadays MOOC learners are increasing globally but the completion rates of chosen courses are very low and dropout rates are very high**[1].** 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%. 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 support online is not possible all the time**[2]**. 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 the large data, including the student ID, years, enrollment date and time, course ID, gender, and the number of courses. This journal is about the data from one university.
By using machine learning techniques we can get more accuracy by deep learning. We also used feature engineering to find the correlation between the features that we took from the dataset**[3,4]**. The data size, processing time, and the chances of the error are very low, and high accuracy for the correlation between the features from the data set by using feature engineering. The dropout rates of the given data set and the accuracy is predicted by some of the Machine

learning techniques such as Random Forest Classifier, Decision tree, Logistic regression, KNN classifier, polynomial SVM, Gradient boost, AdaBoost, MLP Regressor. Finding the accuracy, precision, and ROC Curve for each machine learning algorithm. The goal of our research is all about finding the accuracy and precision of the MOOC course each year and predicting the learners from dropping out of the online courses.

Literature review:

One major issue is the high percentage of dropouts. The majority of 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 or not a student will drop out. The data we have 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.**[5]** In recent years, some techniques were proposed to overcome this problem. The majority 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. Despite the fact that 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. 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.

People that extract features in this procedure must be familiar with the dataset and have some subject knowledge. 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. 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 manually if new types of datasets exist.**[5]** This research focuses on utilizing ML and RF to predict MOOC dropout to predict student dropout. Machine Learning (ML) is a powerful technology that may be used in Learning Analytics to uncover hidden patterns of student involvement in MOOCs, offering advantages above standard statistical analysis. K-Nearest Neighbors (KNN), Decision Trees (DT), RF, and Decision Trees(DT) are examples of machine learning algorithms that have been used to predict students' learning success at the end of a course. We tested in a varied data store, involving official and informal educational aspects, using both statistical and predictive models**[6]**. Students' dropout in an online course was also identified using logistic regression. This approach outperformed logistic regression, Support Vector Machine (SVM), and KNN in terms of precision, recall, specificity, and accuracy during validation. Eight machine learning

algorithms were used to examine students with a high chance of failing in order to detect their performance early.

Random forest, Gradient boost, ADA boost, and MLP Regressor were the most effective algorithms for early student identification. The accuracy was determined to be 95%. Furthermore, data processing was discovered to be crucial in improving the efficiency of ML algorithms[7]. Predictive models have been described in previous studies, however, various constraints limit their use to a single learning platform. It's difficult to create predictive and flexible models that can adapt and/or adjust to varied learning environments. The presence of multiple course structures, instructional designs, and online platforms were all restrictions.

The prediction is made using the Random Forest Model technique in Machine Learning (ML), which is analyzed using validation measures such as accuracy, precision, recall, F1-score, and the Receiver Operating Characteristic (ROC) curve. With an accuracy of 95%, a precision of 96 percent, a recall of 96%, and an F1 score of 96%, the built model can predict whether students will drop out or continue in the MOOC course on any given day[6]. Shapley values were used to explain the contributing features and interactions for the model prediction. Significant features that assist MOOC learners and designers in developing course material, course structure, and delivery were established using a Decision Tree (DT) algorithm. To evaluate the in-course behavior of online students, three MOOC datasets were subjected to a variety of machine learning techniques. According to the authors, the models utilized could be useful in predicting crucial features in order to reduce attrition. These studies aid in the prediction of student outcomes, including dropout rate; however, none of them forecast students who are in danger of dropping out at a later point in the course. As a result, we present the RF model with features such as including average, standard deviation, variance, and skewness[6]. The literature review presented in this article is primarily divided into two categories: methodologies and objectives. The strategies are described first in this article since they are used to achieve the goals discussed in each reference. These strategies are then implemented using a variety of algorithmic ways.

Objectives: The objectives are linked to the students' learning processes' interests and concerns. Techniques: The techniques take into consideration various algorithms, methods, and tools that process data in order to assess and anticipate the sign of things to come[7]. The popular machine learning approaches, namely, support vector machines (SVM), and, were utilized in this study to predict student dropouts. These strategies have been successfully applied to a variety of classification issues, and they operate in two phases: training and testing. During the training phase, each approach is given a set of sample data pairs (X, Y), where X represents the pair's input and Y represents the pair's output. In this study, Y can be either 0 or 1 if a student completes the course or 1 if the student drops the course. As a result, predicting dropouts is a two-class classification problem[8]. The internal parameters of each technique are then adjusted to infer the mapping implied by the data. During the testing phase, each technique is given data

that was not utilized during the training phase in order to evaluate its classification performance. If the machine learning technique is found to correctly identify the majority of the data in the test set, the training is regarded as successful, and the machine learning technique displays generalization capabilities. Because most issues do not have linearly separable data, the above technique is usually modified to handle non-linearly separable data at the risk of adopting a number of misclassifications[9]. The SVM technique turns the data into a feature space with a greater dimension than the input before intending to separate them using a linear discriminator to improve its performance. It accomplishes this with the help of a kernel function. There are many different types of kernel functions, including linear, polynomial, and radial basis function kernels. The SVM technique translates the data from the testing sample into the feature space that it utilized for training and then classifies it during the testing phase.

Dataset Description:

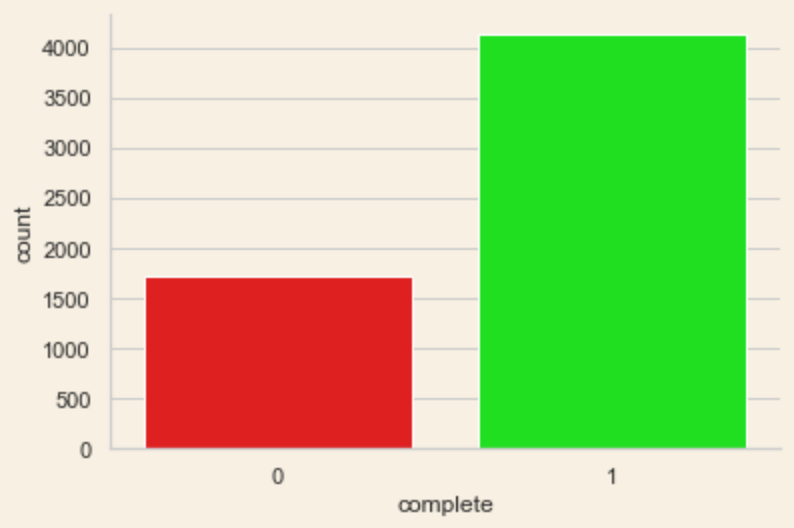
Prediction of early dropouts from online courses conducted by the Student course dropout prediction dataset are included in the collection of data like student_id, years, entrance_test_score, enroll_date_time, course_id, Course_progress, online, gender, international, complete[10]. From 2016 to 2020, 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. The availability of this data is restrictive. Data has been collected through EdPlus and is accessible with permission from the authors. EdPlus is a service that is provided by EdPlus. The family educational and privacy act protects student information (FERPA).

Proposed Methodology:

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 shows the distribution of students in this course[10]. The student demographic data were evaluated to gain a sense of the students' backgrounds. students and such a description benefit us in comprehending the significance of this study. **Table** shows the distribution of students in this course.

Table:

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

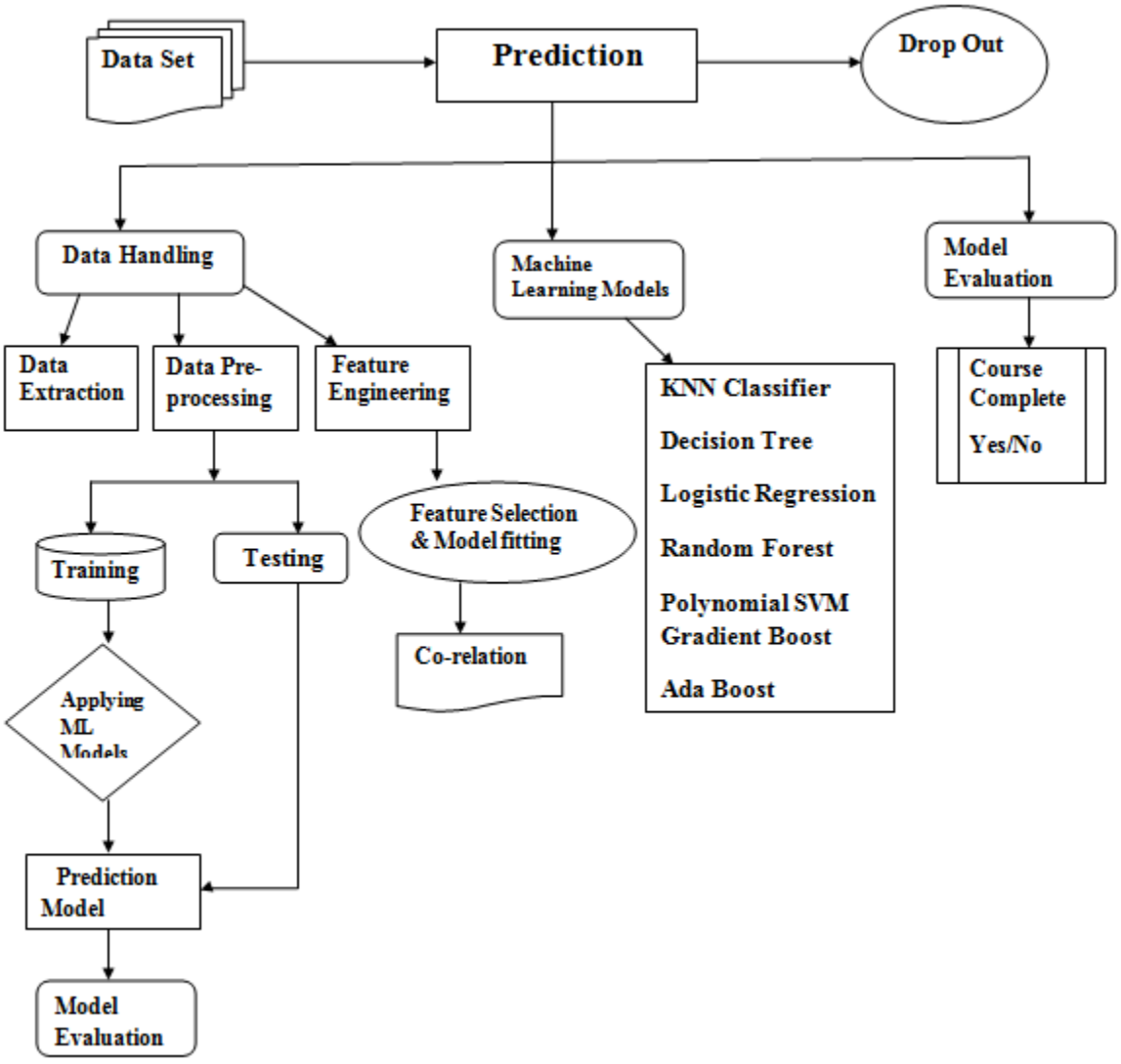


Distribution of Students in Courses

Here, From the ASU Dataset, the Completion, of course, is high when compared to dropouts from the courses.

Data analysis, machine learning modeling, and model assessment are the three aspects of the process. The techniques' flow is represented in a flowchart.

Flow Chart



The above flow chart demonstrates how we categorize prediction techniques. Three distinct classifications have been identified: Data handling, Machine learning Models, and Model evaluation. Here initially we had imported/Extracted the ASU data set and we need to do preprocessing like we need to check the null values in the data set if we have the null values in our data set we need to remove all the null values in our data set to increase the rate of accuracy for our models. After cleaning the data we need to divide the data set to train and test our machine learning model, so that 80% of the data is required to train the model and 20% of data is sufficient to test the model. Next is our feature engineering here we

Data Handling: The data is handled according to the Arizona State University dataset's standard procedure. Three phases are involved in data handling: data extraction, data preprocessing, data cleaning, and feature engineering.

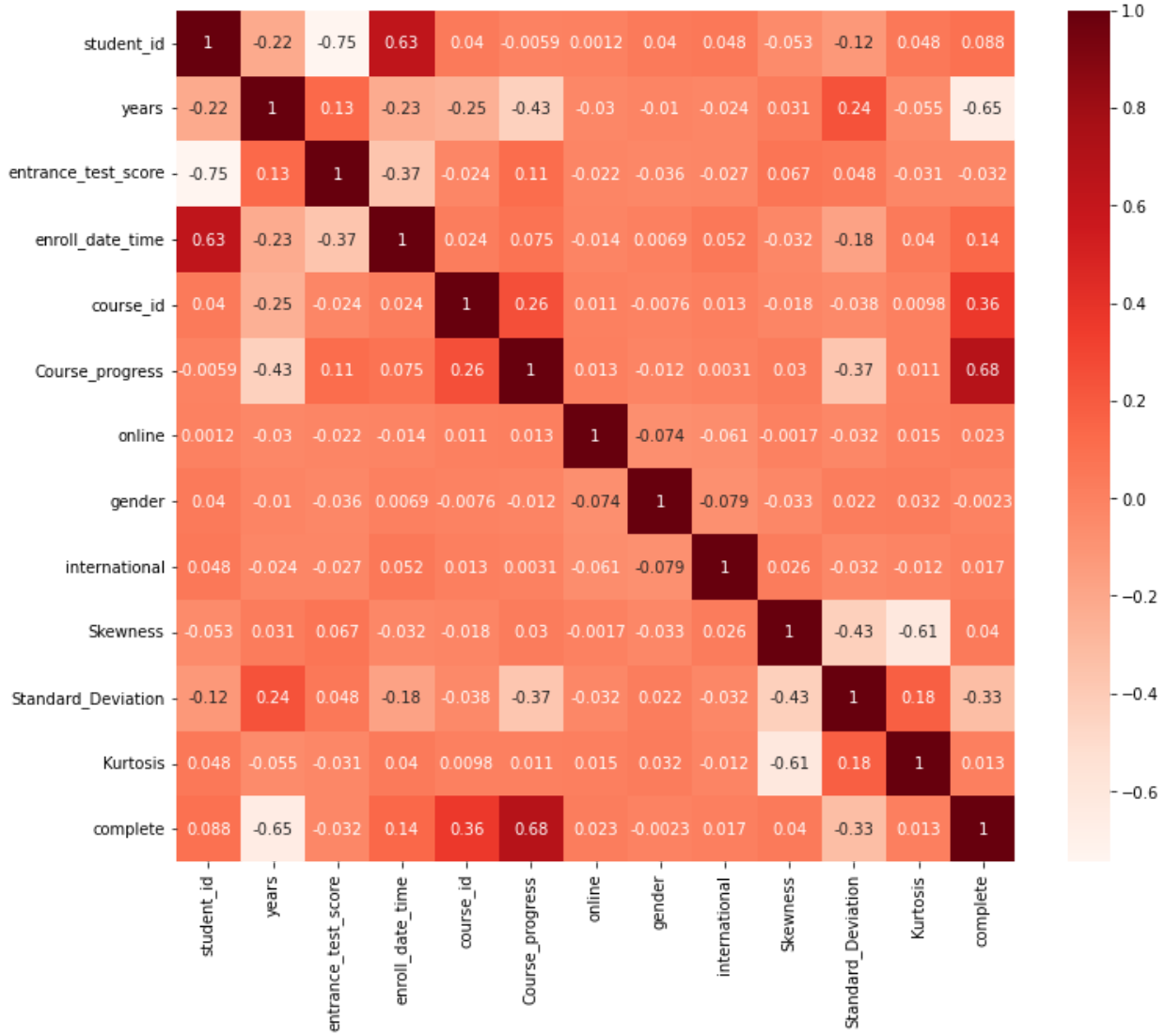
Data Extraction: Data extraction is the process of obtaining information from a data source using data mining techniques. The ALEKS platform has an API for accessing the data in the SQL Database. The data selection technique was used to choose a query from the data source **[11]** using the data selection method. SQL is a database. The query table is saved as a database after the data selection is completed. CSV file (comma-separated values) The CSV file was accessed using a Jupyter notebook. The python programming language is used to process the data. This course has a total of 5861 students, all of whom have some degree of proficiency. Following their Initial Knowledge Check, they will engage in an activity (IKC). 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 and Data Cleaning:

Data preprocessing is the procedure for processing original data to be used in a machine learning model. It's the most crucial step in creating a machine learning model. Data preprocessing is a critical step in cleaning and preparing data for a machine learning model, which enhances the model's accuracy and efficiency. The characteristics of student learning change that are linked to student dropout have been identified and utilized to predict this event. The data in the extracted data set is a time series with no primary keys. The raw dataset was organized using student identification as a criterion for grouping **[11]**. After the target data has been aggregated, the learning progression data for each student may be retrieved and placed in a table called the preprocessed data.

The panda's library is a data cleaning and analysis tool. 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 for teaching. This covers classification, regression, clustering, and dimensional reduction, among other useful techniques for machine learning and statistical modeling. Each x represents the sequence of subjects mastered by the student from the beginning of the course to the day in question, and each y is the label for that sequence, suggesting if the student has completed the course. On that day, students either remained in the course or dropped out. This results in the data that has been preprocessed in

preparation for the machine learning model most of the preprocessed characteristics, Because the data is of several data kinds, it is cleaned to be of the same numerical data type of modeling.



Feature Engineering

The data is turned into a feature table, with each row representing the rate of student performance on each given day since the course began. In addition, a target column is added to the table, which displays the label of the student's learning for that row. If that row of learning resulted in a student dropping out of a course, the target column was tagged with a "1." If a learner had another day of learning following that day, that day was designated as "0." indicating that the student stayed in the class.

Student's learning Graph shows how topics are learned over time.

The average, standard deviation, variance, skewness, and kurtosis are the most commonly used statistical characteristics [12,13]. The above illustrates a graph of student learning expressed as topic knowledge over time. Because the curve is so skewed, the average cannot accurately represent the pace of learning. As a result, the time-series average computed in windows was achieved[16]. The moving average is a list of averages that, when combined with the average, can provide an overview of the rate of learning. The normalized value of this list is utilized as a feature. This provides four characteristics that may be used to indicate the process of progress in learning[14]. Because the curve in Student's learning Graph shows how topics are learned over time is quite rough, four attributes are used to represent it: skew, standard deviation, variance, and kurtosis[21,20]. Topics mastered, overall trajectory, and ultimate trajectory is the additional characteristics considered in the study. The total trajectory is determined as the connection between the first and last day of distribution.

It's difficult to develop predictive and flexible models that can adapt and/or adjust to varied learning settings. The prevalence of multiple course formats, educational designs, and online platforms [15] are all limitations. Machine Learning (ML) is an advanced tech that may be used in Learning Analytics to uncover hidden patterns of student interaction in MOOCs [4,18,19]. It has several benefits over traditional statistical analysis. K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Trees (DT), RF, and Deep Learning (DL), are examples of machine learning algorithms that have been used to predict students' learning achievement at the conclusion of a course [21].

Modeling and Machine Learning

The data for the machine learning model is stored in the feature table after it is constructed. The Machine Learning modeling predicts MOOC dropout based on the supplied input characteristics[16,17]. 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 K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Trees (DT), RF, Deep [9]Learning (DL), and Recurrent Neural Network (RNN), have been applied to the prediction of students learning performance at the end of the course [22]. Out eight 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, MLP Regression, and AdaBoost Classifier algorithms were the most successful at identifying learners immediately. The accuracy was determined to be 96%[10]. Furthermore, data cleaning has been proven to be helpful in improving the performance of machine learning algorithms [23]. These studies improve in the prediction of academic achievements, including dropout rate; however, none of them estimates students who are in danger of dropping out at a later point in the course. Furthermore, no research has been done on the use of RF to predict student dropout using the characteristics revealed in this study. As a result, we present the RF model with average, standard deviation, variance, skew, kurtosis, moving average, overall trajectory, and ultimate trajectory as well as other properties.

Model Fitting and Feature Selection

The produced characteristics were analyzed and approved in this stage[4]. The correlation matrix approach of Exploratory Data Analysis (EDA) was utilized to assess the characteristics in order to predict student learning outcomes in MOOCs [25]. As a result, data leakage in the ML model would be avoided, and the modeling project's success would be increased.

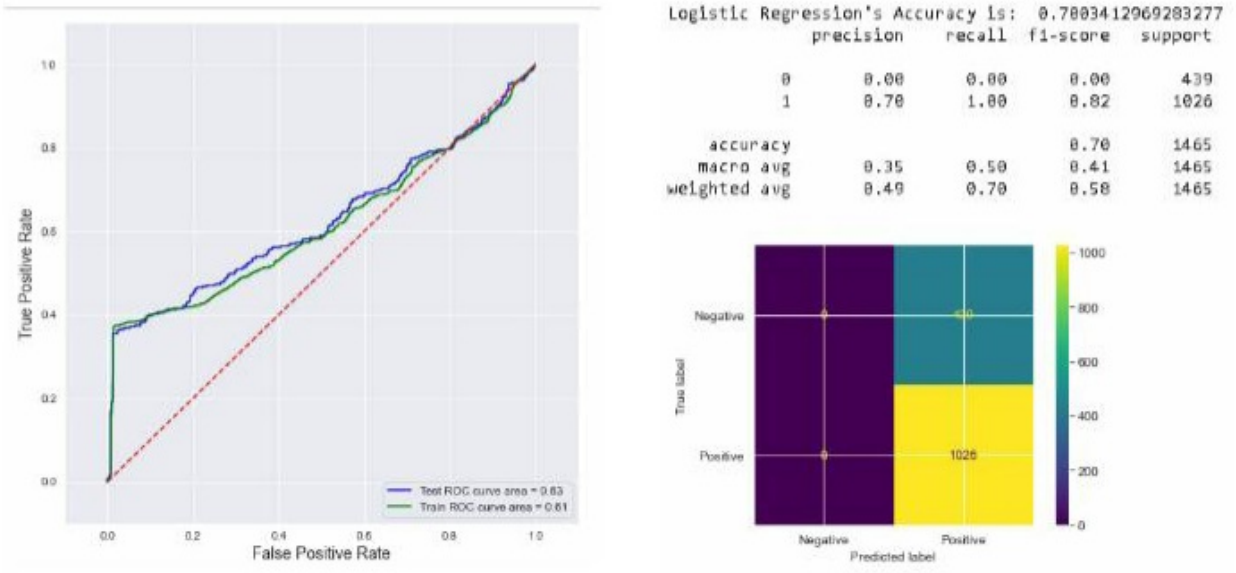
Model Prediction for Testing and Training

Despite going through feature engineering and feature selection methods, the dataset in this study still contains a significant number of outliers, hence random forest is a better fit for this study. Models based on machine learning These models were used to train with the training data and were then tested. ability to forecast if a student will fail the course or complete it when the model was fed information on a student's learning in a MOOC[10]. 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[21]. 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 after the data was balanced. 75%of the produced data was utilized to train the model, while the remaining 25% was used to test the model.

Results and Discussions

Course drop out(0) and complete(1) students Performances of Different Methods under Different Metrics shows the experimental outcomes of our model and baseline approaches. As can be seen from the table, when compared to standard classification techniques Our model can produce equivalent outcomes on all of the problems (engineering).

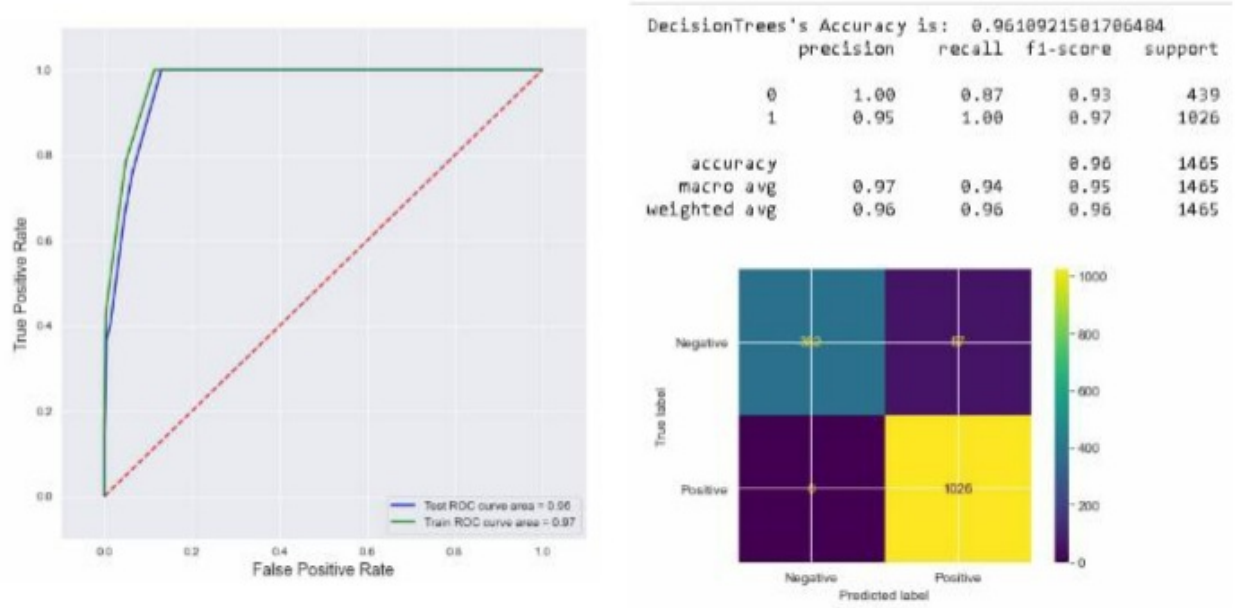
Logistic regression: In an online learning course, logistic regression was also used to predict student dropout [23]. Precision, recall, and accuracy were all validated using this approach. To accomplish dropout prediction based on series classification, logistic regression (LR), statistical analysis, and other methods were used**[10]**. Depending on whether the categorization is binary or a temporal series, When it comes to classification problems, the classifier has a role in predicting dropouts**[28,29]**. 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.



Roc Curve And Confusion Matric 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 [23]. Because of its great accuracy, the Decision Tree was selected as the best appropriate algorithm. A decision tree is a map (tree-like structure) that shows the various outcomes of a decision. The Decision Tree algorithm uses actual values to create a tree. The Decision Tree is created using the divide-and-conquer strategy. A cluster of the Decision Tree

may be recognized in the Random Forest classification technique **[30][22].** The Decision Tree method was chosen as the most suited because of its extreme accuracy. A decision tree is a diagram that shows the many outcomes of a selection**[31**

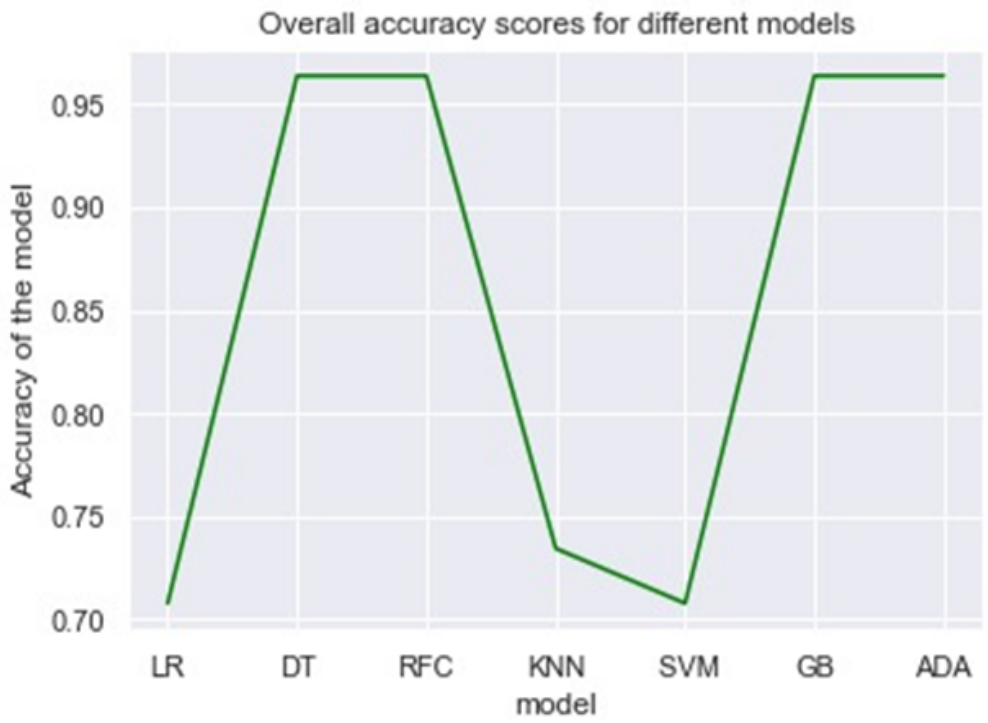


Roc Curve And Confusion Matric for Decision tree Model

Model	Accuracy	Precision		F1 Score		Recall		ROC Curve	
		0	1	0	1	0	1	Test	Train
Logistic Regression	0.700	0.00	0.70	0.00	0.82	0.00	1.00	0.62	0.62
Decision Tree	0.961	1.00	0.95	0.93	0.97	0.87	1.00	0.97	0.97
Random Forest	0.961	1.00	0.95	0.93	0.97	0.87	1.00	0.87	1.00
KNN Classifier	0.731	0.57	0.78	0.48	0.86	0.42	0.86	0.77	0.85
Polynomial SVM	0.700	0.00	0.70	0.00	0.82	0.00	1.00	0.72	0.71
Gradient Boost	0.961	0.57	0.78	0.48	0.82	0.42	0.86	0.96	0.97
ADA Boost	0.959	1.00	0.95	0.93	0.97	0.87	1.00	0.97	0.98

Course drop out(0) and Course complete(1) students
Performances of Different Methods under Different Metrics

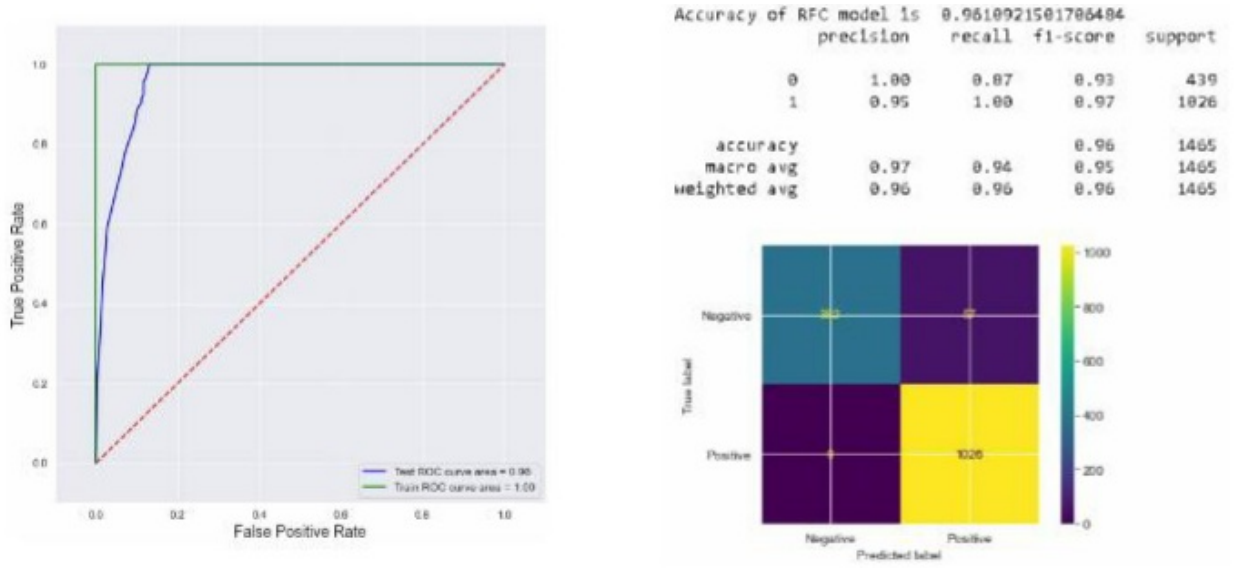
This table indicates the machine learning models which we used in our project with different rates of accuracy, precision, F1 score, Recall, and ROC Curve for training and testing.



Accuracy Score for Different ML Models

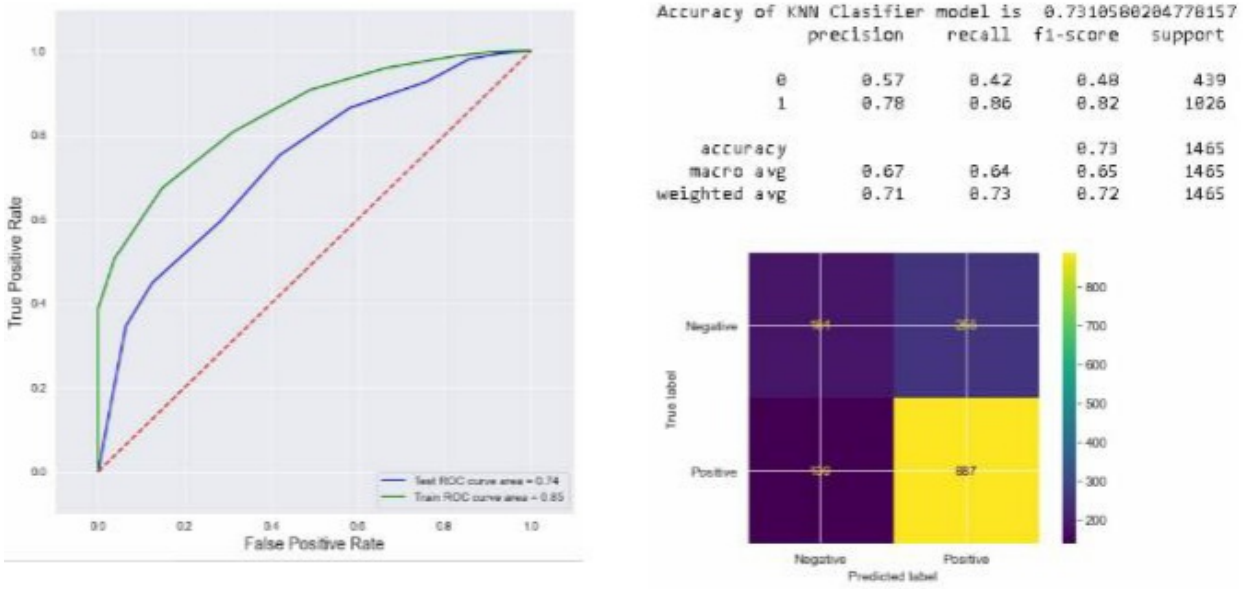
This is the graph of the overall performance of accuracy scores for different Machine Learning Models

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**[34]**. The Decision Tree cluster is visible in the Random Forest classification technique. 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, with the Decision Tree method being chosen for analyses because of its comparatively high accuracy among all the ml methods. The RF model's accuracy is on a level with, if not greater than, that of most ML models**[26,27]**. Outliers and noise are less noticeable in RF**[10]**. In this study, the internal estimates of error, strength, correlation, and variable relevance are really valuable. 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**[35]**. As a result, RF is particularly reliable when predicting a noisy dataset with a lot of outliers. Although the dataset in this study has gone through thorough feature engineering and feature selection processes, there are still a lot of exceptions, hence random forest is the right match for this study.

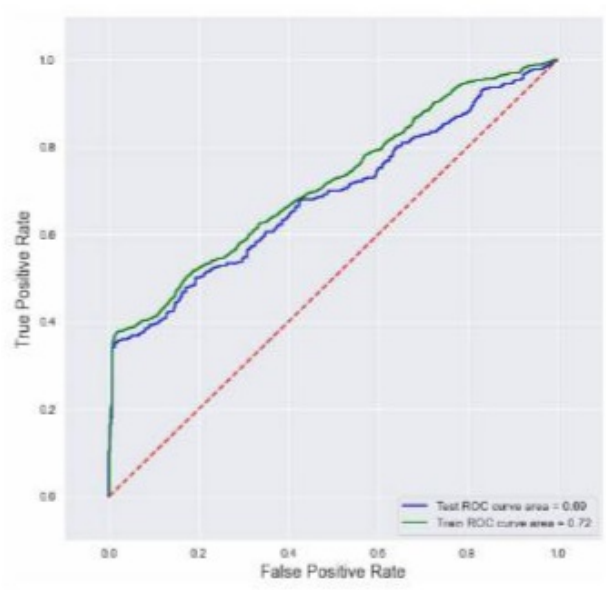


Roc Curve And Confusion Matric for Random Forest Model

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 **[32]** 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),**[33]** train an SVM classifier based on the RBF(radial basis function) kernel. and case models in general). The writers differentiate between cases of dropouts and passive students. Dropouts are a group of academics that do not connect with one another at all. Inactive learners continue to attend classes.

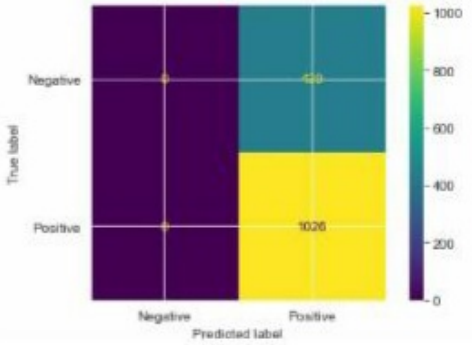


Roc Curve And Confusion Matric for K-Nearest Neighbour Model



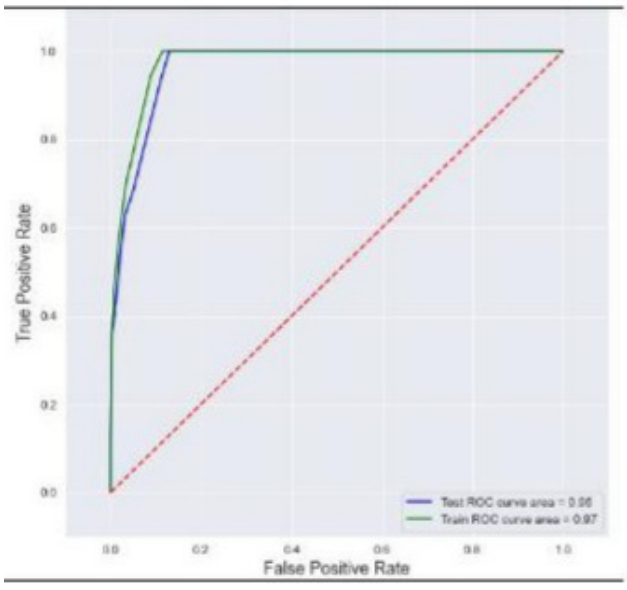
Accuracy of SVM model is 0.7883412969283277

	precision	recall	f1-score	support
0	0.00	0.00	0.00	439
1	0.70	1.00	0.82	1026
accuracy			0.70	1465
macro avg	0.35	0.50	0.41	1465
weighted avg	0.49	0.70	0.58	1465



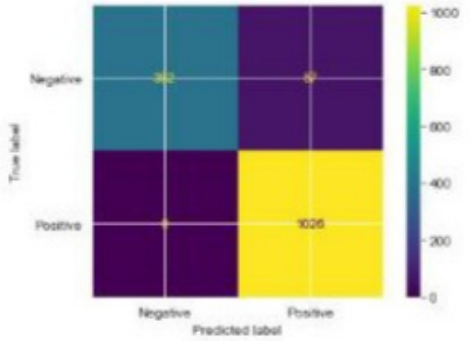
Roc Curve And Confusion Matric for SVM Model

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 **[37]** and AdaBoost **[38]** are examples of ensemble algorithms in the SDP literature. To find the top-performing variation for their dataset, we used the Tri-Training **[40]** method. Easy-to-interpret approaches such as decision trees are pitted against noninterpretable strategies **[39]**. (e.g., random forests and gradient-boosted trees).

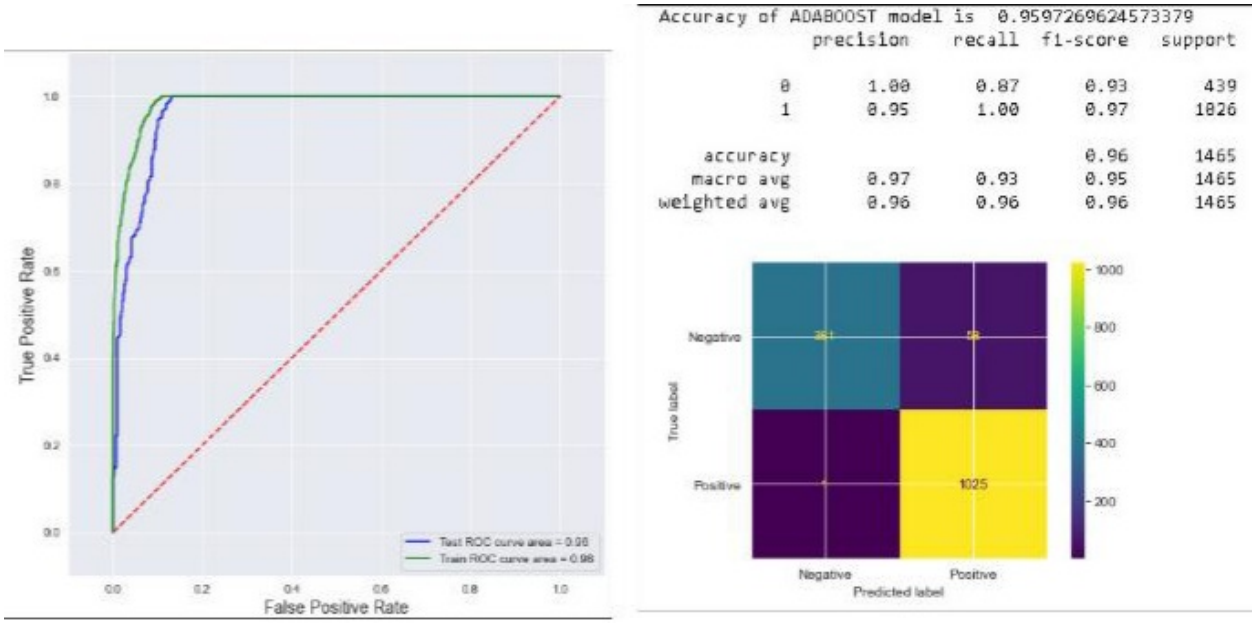


Accuracy of GB model is 0.9610921501706484

	precision	recall	f1-score	support
0	0.57	0.42	0.48	439
1	0.78	0.86	0.82	1026
accuracy			0.73	1465
macro avg	0.67	0.64	0.65	1465
weighted avg	0.71	0.73	0.72	1465



Roc Curve And Confusion Matric for Gradient Boost Model



Roc Curve And Confusion Matric for Ada Boost Model

Future Scope

DPM's (dropout prediction model) machine learning study has sparked the interest of many scholars. The estimation and execution techniques of the initial mass for students are first offered as per Ml Model, and also suggested the training's furthest zone defining sample, as well as modification and execution[27]. We initially 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[36]. In the DPM study based on machine learning, several obstacles must be solved, such as learner distance, training sample quality, training sample imbalance, feature extraction and representation, and so on. We used eight algorithms to build the machine learning model in this project. As a consequence, we will be able to develop a variety of models in the future utilizing a variety of machine learning approaches and data properties[10].

Conclusion

The problem of predicting whether students would drop out from 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've made a first step in detecting at-risk students early and correctly, which will help academics in developing interventions. We several prediction models and found that regularised decision tree, Random

Forest, and Gradient Boost performed best in terms of accuracy. Despite the fact that 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). Precision, Recall, F1-score, AUC, and ROC curve are used to evaluate features and modeling in the dataset, With an accuracy of 0.96, precision of 0.94, recall of 0.97, F1-score of 1.00, and an AUC of 0.96, this can predict student dropout to an acceptable standard in the research community.

References

[1]. Herbert Schildt, Java Complete Reference, 5th edition- Dietel and Dietel, Java How To program- Pressman, Software Engineering, 4thedition

[2]. Rahul Katarya;Jalaj Gaba;Aryan Garg;Varsha Verma; (2021). A review on machine learning based student's academic performance prediction systems . 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), (), –.

[3]. E. S. Bhutto, I. F. Siddiqui, Q. A. Arain and M. Anwar, "Predicting Students' Academic Performance Through Supervised Machine Learning," *2020 International Conference on Information Science and Communication Technology (ICISCT)*, 2020, pp. 1-6, doi: 10.1109/ICISCT49550.2020.9080033.

[4] John Daniel. 2012. Making sense of MOOCs: Musings in a maze of myth, paradox and possibility. Journal of interactive Media in education 2012, 3 (2012).

[5]. Wei Wang, Han Yu, and Chunyan Miao. 2017. Deep Model for Dropout Prediction in MOOCs. In Proceedings of ICCSE'17, Beijing, China, July 6–9, 2017, pages .https://doi.org/10.1145/3126973.3126990

[6]. Dass, S.; Gary, K.; Cunningham, J. Predicting Student Dropout in Self-Paced MOOC Course Using Random Forest Model. Information 2021, 12, 476.

[7]. Ioanna Lykourentzou *, Ioannis Giannoukos, Vassilis Nikolopoulos, George Mpardis, Vassili Loumo. Lykourentzos et al. / Computers & Education 53 (2009) 950–965

[8] Fei Mi and Dit-Yan Yeung. 2015. Temporal models for predicting student dropout in massive open online courses. In 2015 IEEE International Conference on Data Mining Workshop (ICDMW). IEEE, 256–263.

[9] Tanmay Sinha, Patrick Jermann, Nan Li, and Pierre Dillenbourg. 2014. Your click decides your fate: Inferring information processing and attrition behavior from mooc video clickstream interactions. In 2014 Empirical Methods in Natural Language Processing Workshop on Modeling Large Scale Social Interaction in Massively Open Online Courses.

[10] Dass, Sheran, Kevin Gary, and James Cunningham. 2021. "Predicting Student Dropout in Self-Paced MOOC Course Using Random Forest Model" *Information* 12, no. 11: 476.

[6]. Marius Kloft, Felix Stiehler, Zhilin Zheng, and Niels Pinkwart. 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. 60–65.

[8] Cameron C. Gray and Dave Perkins. 2019. Utilizing early engagement and machine learning to predict student outcomes. Comp. & Education 131 (2019).

[9] Jiazhen He, James Bailey, Benjamin I.P. Rubinstein, and Rui Zhang. 2015. Identifying at-risk students in massive open online courses. In the Twenty-Ninth AAAI Conference on Artificial Intelligence.

[10] Ya-Han Hu, Chia-Lun Lo, and Sheng-Pao Shih. 2014. Developing early warning systems to predict students' online learning performance. Computers in Human Behavior 36 (2014), 469–478

[11]. Saa, A.A. Educational Data Mining & Students' Performance Prediction. Int. J. Adv. Comput. Sci. Appl. 2016, 7, 212–220.

[12]. Nanopoulos, A.; Alcock, R.; Manolopoulos, Y. Feature-based classification of time-series data. Int. J. Comput. Res. 2001, 10, 49–61.

[13]. Doanne, D.; Seward, L.E. Measuring skewness. J. Stat. 2011, 19, 1–18.

[14] Gerben W. Dekker, Mykola Pechenizkiy, and Jan M. Vleeshouwer. 2009. Predicting Students Drop Out: A Case Study. International Working Group on Educational Data Mining (2009).

[15]. Gaševi´c, D.; Dawson, S.; Rogers, T.; Gasevic, D. Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. Internet High. Educ. 2016, 28, 68–84. [CrossRef]

[16]. Qiu, J.; Tang, J.; Liu, T.X.; Gong, J.; Zhang, C.; Zhang, Q.; Xue, Y. Modeling and Predicting Learning Behavior in MOOCs. In Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, CA, USA, 22–25 February 2016; pp. 93–102. [CrossRef]

[17]. Baker, R.S.J.D.; Siemens, G. Educational Data Mining and Learning Analytics. In Cambridge Handbook of the Learning Sciences, 2nd ed.; Keith Sawyer, R., Ed.; Cambridge University Press: New York, NY, USA, 2014; pp. 253–274.

[18]. Al-Shabandar, R.; Hussain, A.; Laws, A.; Keight, R.; Lunn, J.; Radi, N. Machine learning approaches to predict learning outcomes in Massive open online courses. In Proceedings of the International Joint Conference on Neural Networks, Anchorage, AK, USA, 14–19 May 2017; pp. 713–720. [CrossRef].

[19]. Gaševi´c, D.; Rose, C.; Siemens, G.; Wolff, A.; Zdrahal, Z. Learning Analytics and Machine Learning. In Proceedings of the Fourth International Conference on Learning Analytics and Knowledge, Indianapolis, IN, USA, 24–28 March 2014; pp. 287–288. [CrossRef]

[20]. Bozkurt, A.; Yazıcı, M.; Aydin, I.E. Cultural diversity and its implications in online networked learning spaces In Research Anthology on Developing Effective Online Learning Courses; Information Resources Management Association, Ed.; IGI Global: Hershey, PA, USA, 2018; pp. 56–81.

[21]. Baker, R.S.; Inventado, P.S. Educational Data Mining and Learning Analytics in Learning Analytics; Springer: New York, NY, USA, 2014; pp. 61–75.

[22]. K″orösi, G.; Farkas, R. Mooc performance prediction by deep learning from raw clickstream data. In Proceedings of the International Conference in Advances in Computing and Data Sciences, Valletta, Malta, 24–25 April 2020; pp. 474–485.

[23]. Adnan, M.; Habib, A.; Ashraf, J.; Mussadiq, S.; Raza, A.A.; Abid, M.; Bashir, M.; Khan, S.U. Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models. IEEE Access 2021, 9, 7519–7539. [CrossRef]

[24]. Burgos, C.; Campanario, M.L.; De La Peña, D.; Lara, J.A.; Lizcano, D.; Martínez, M.A. Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. Comput. Electr. Eng. 2018, 66, 541–556. [CrossRef]

[25]. Al-Shabandar, R.; Hussain, A.; Laws, A.; Keight, R.; Lunn, J.; Radi, N. Machine learning approaches to predict learning outcomes in Massive open online courses. In Proceedings of the International Joint Conference on Neural Networks, Anchorage, AK, USA, 14–19 May 2017; pp. 713–720. [CrossRef]

[26] Alamri, A.; Alshehri, M.; Cristea, A.; Pereira, F.D.; Oliveira, E.; Shi, L.; Stewart, C. Predicting MOOCs Dropout Using Only Two Easily Obtainable Features from the First Week's Activities. Intelligent Tutoring Systems; Coy, A., Hayashi, Y., Chang, M., Eds.; Springer: Cham, Switzerland, 2019; pp. 163–173. [CrossRef]

[27.]Breiman, L. Random forests. Mach. Learn. 2001, 45, 5–32. [CrossRef]

[28] Jiazhen He, James Bailey, Benjamin I.P. Rubinstein, and Rui Zhang. 2015. Identifying at-risk students in massive open online courses. In Twenty-Ninth AAAI Conference on Artificial Intelligence.

[29] Sreerama K Murthy. 1998. Automatic construction of decision trees from data: A multi-disciplinary survey. Data mining and knowledge discovery 2, 4 (1998), 345–389

[30] Qasem A. Al-Radaideh, Emad M. Al-Shawakfa, and Mustafa I. Al-Najjar. 2006. Mining student data using decision
trees. In Proceedings of the International Arab Conference on Information Technology (ACIT'2006). 1–5.

[31] Jing Chen, Jun Feng, Xia Sun, Nannan Wu, Zhengzheng Yang, and Sushing Chen. 2019. MOOC Dropout Prediction Using a Hybrid Algorithm Based on Decision Tree and Extreme Learning Machine. Mathematical Problems in Engineering (2019)

[32] Marius Kloft, Felix Stiehler, Zhilin Zheng, and Niels Pinkwart. 2014. Predicting MOOC dropout over weeks using machine learning methods. In Proceedings of the EMNLP Workshop on Analysis of Large Scale Social Interaction in MOOCs. 60–65

[33] Bussaba Amnueypornsakul, Suma Bhat, and Phakpoom Chinprutthiwong. 2014. Predicting attrition along the way: The UIUC model. In Proceedings of the EMNLP Workshop on Analysis of Large Scale Social Interaction in MOOCs. Association for Computational Linguistics, 55–59

[34]. Breiman, L. Random forests. Mach. Learn. 2001, 45, 5–32. [CrossRef]

[35] Archer, K.J.; Kimes, R.V. Empirical characterization of random forest variable importance measures. Comput. Stat. Data Anal. 2008, 52, 2249–2260. [CrossRef]

[36]. Alamri, A.; Alshehri, M.; Cristea, A.; Pereira, F.D.; Oliveira, E.; Shi, L.; Stewart, C. Predicting MOOCs Dropout Using Only Two Easily Obtainable Features from the First Week's Activities. Intelligent Tutoring Systems; Coy, A., Hayashi, Y., Chang, M., Eds.; Springer: Cham, Switzerland, 2019; pp. 163–173. [CrossRef]

[37] Cameron C. Gray and Dave Perkins. 2019. Utilizing early engagement and machine learning to predict student outcomes. Comp. & Education 131 (2019).

[38] Ya-Han Hu, Chia-Lun Lo, and Sheng-Pao Shih. 2014. Developing early warning systems to predict students' online learning performance. Computers in Human Behavior 36 (2014), 469–478.

[39]. Eranki, K.L.; Moudgalya, K.M. Evaluation of web based behavioral interventions using spoken tutorials. In Proceedings of the 2012 IEEE Fourth International Conference on Technology for Education, Hyderabad, India, 18–20 July 2012; pp. 38–45.

[40]. Wang, J.; Xu, M.; Wang, H.; Zhang, J. Classification of imbalanced data by using the SMOTE algorithm and locally linear embedding. In Proceedings of the 2006 8th International Conference on Signal Processing, Guilin, China, 16–20 November 2006; Volume 3.