9-29-2023

Predictive Analysis of Students' Learning Performance Using Data Mining Techniques: A Comparative Study of Feature Selection Methods

S. M. F. D. Syed Mustapha
Zayed University

applied system innovation

MDPI

Article

Predictive Analysis of Students' Learning Performance Using Data Mining Techniques: A Comparative Study of Feature Selection Methods

S. M. F. D. Syed Mustapha

College of Technological Innovation, Zayed University, Dubai P.O. Box 144534, United Arab Emirates; syed.duani@zu.ac.ae

Abstract: The utilization of data mining techniques for the prompt prediction of academic success has gained significant importance in the current era. There is an increasing interest in utilizing these methodologies to forecast the academic performance of students, thereby facilitating educators to intervene and furnish suitable assistance when required. The purpose of this study was to determine the optimal methods for feature engineering and selection in the context of regression and classification tasks. This study compared the Boruta algorithm and Lasso regression for regression, and Recursive Feature Elimination (RFE) and Random Forest Importance (RFI) for classification. According to the findings, Gradient Boost for the regression part of this study had the least Mean Absolute Error (MAE) and Root-Mean-Square Error (RMSE) of 12.93 and 18.28, respectively, in the case of the Boruta selection method. In contrast, RFI was found to be the superior classification method, yielding an accuracy rate of 78% in the classification part. This research emphasized the significance of employing appropriate feature engineering and selection methodologies to enhance the efficacy of machine learning algorithms. Using a diverse set of machine learning techniques, this study analyzed the OULA dataset, focusing on both feature engineering and selection. Our approach was to systematically compare the performance of different models, leading to insights about the most effective strategies for predicting student success.

Keywords: data mining; feature selection methods; Boruta algorithm; lasso regression; recursive feature elimination (RFE); random forest importance (RFI)

Citation: Syed Mustapha, S.M.F.D. Predictive Analysis of Students' Learning Performance Using Data Mining Techniques: A Comparative Study of Feature Selection Methods. *Appl. Syst. Innov.* **2023**, *6*, 86. https://doi.org/10.3390/asi6050086

Academic Editor: Patrícia Ramos

Received: 6 July 2023
Revised: 2 September 2023
Accepted: 18 September 2023
Published: 29 September 2023

1. Introduction

In the modern educational landscape, the ability to predict and understand student performance is paramount. It not only provides educators with valuable insights into learning patterns but also allows students to recognize areas of improvement, fostering a conducive learning environment. This research delves into the application of machine learning (ML) techniques for predicting students' academic performance, addressing a critical gap in contemporary educational research.

Numerous scholarly investigations have examined diverse aspects of student performance. However, the swift improvements in machine learning present a novel viewpoint, holding the potential to provide more precise and nuanced forecasts. The current research, based on an empirical framework, aims to utilize these improvements, with a specific emphasis on evaluating the effectiveness of various machine learning models in predicting academic outcomes. This research endeavors to illuminate the most effective ways for academic forecasting by contrasting classic prediction methods with state-of-the-art machine learning models. The next sections explore the methodology employed, the datasets utilized, and the insights obtained, resulting in a full comprehension of the role and promise of machine learning in forecasting student outcomes.

Students' performance prediction and students' academics analytics are the two most demanding research topics in the domain of educational literature. Despite having diverse

goals, performance analysis has a substantial impact on prediction research. Learning analytics, according to the common definition, is the process of collecting and analyzing information on students and their surroundings with the goal of improving both [1]. The field of educational data mining is closely related to the domain of learning analytics. Educational data mining, in turn, is concerned with the application of data mining, machine learning (ML), and the statistics of information generated and gathered within educational settings for discovering knowledge about how people learn. Learning analytics is a field that focuses on the application of data mining, ML, and statistics to information generated and gathered within educational settings [2]. It is possible that statistical techniques are insufficient for reliably linking individual variables to outcomes [3]. Supervisors and learners alike may benefit from the groundbreaking insights that may be gleaned by using complex algorithms [4,5]. Many investigators have been motivated to dig further into the information sharing procedure through the rapid development of data mining tools. A few of them have been using data mining techniques for this goal for quite some time [6]. However, at the present moment, there are very few of these kinds of investigations. This effort to enhance the quality of learning is facilitated by the increasing accessibility of digital data gathered by various academic resource management platforms and educational tools in recent years [7–9].

1.1. Learning Management System and Students' Academics Analytics

Nowadays, Learning Management Systems (LMSs) and Open Course Ware (OCW) platforms are a playing a challenging role for the application and exploitation of learning analytics because they are designed to host educational materials (open and web-based materials in the case of OCW platforms), typically organized as courses. Open Course Ware (OCW) platforms, for example, offer adaptable learning environments and enable the automated gathering of learning data like student activity and performance, while also including course preparation materials and assessment tools. Thus, when properly mined using data-mining methods, this massive trove of educational data might be considered as a source of essential information for educators at all levels. Data, mostly from online education environments, that may be utilized to enhance students' learning are made accessible because of the widespread adoption of digital equipment in the classroom. Commonly, learning analytics have been used in a wide variety of settings and organizations [10,11].

Learning analytics may be interpreted and used in a number of ways, including as a tool for analyzing students' performance and habits in comparison to those of their peers. To be more specific, learning analytics may be used in a number of contexts, including easing the delivery of specific indicators for student progress, enabling the online customization of course modules automatically, etc. Predicting how well a student will do on an exam or other kind of assessment is one of the most often used and beneficial uses of learning analytics. This is thought to be especially useful for identifying students "at danger" of dropping out or failing a course in order to offer them with extra help or tutoring in a timely manner (e.g., to identify students who are likely to fail an end-of-semester test). This is particularly crucial when a large number of students are enrolled in a course offered through distant learning. More than that, it might be put to use in providing individualized learning plans and assessments to students according to their prior performance and areas of interest. Teachers could benefit from the data gleaned from these learning analytics by learning which courses and teaching programs need improvement, adaptation, or development of new curriculum offerings, etc., and then hiring or planning interventions to better support students (either individually or in groups).

There has been a lot of research published on the topic of student test performance analytics. Students' test performance was the primary focus of these investigations, and the categorization challenge was solved by dividing them into "pass" and "fail" groups. The goal of these analyses was to identify which students were most likely to fail a certain class. Kotsiantis et al. [12] used machine learning methods like Naive Bayes (NB) and k-Nearest Neighbors (kNN) to classify students as dropouts or not. Findings from the

research revealed that students at risk of dropping out may be identified by looking at merely their demographic information or, more succinctly, their background. Possible EDM goals were outlined and research reported between 1995 and 2005 was consolidated [13]. More recently, in 2010, they [14] released yet another comprehensive assessment that looked at approximately 300 studies from 1993 to 2009. This framework categorizes research on EDM into eleven distinct categories, one of which is the forecasting of student achievement. By that time, learning analytics as a field had also taken off. The analytical investigations in education are first driven by two research communities, the worldwide academic data mining society as well as the society for educational analytics research, predicting students' final exam marks in an online class [15]. The authors of this study evaluated three distinct ways of classifying students' performance: categorizing them into two classes ("pass" and "fail"); three classes ("high", "middle", and "poor"); and nine classes.

Despite EDM approaches having been used by academics in both conventional and computer-based online education, the use of the former is somewhat less common than the latter. In light of this, the researchers of two EDM review publications [14,16] discovered very little research focusing on the conventional schooling model. Furthermore, all individual EDM polls of student achievement takes into account the findings on both online and conventional classroom-based learning. It is worth noting that authors looked high and low for a survey paper on EDM that specifically addressed education in the classroom but came up empty. Predictors, methodologies, and prediction efficiency have been the primary foci of prior surveys of student performance. No one, however, considers the passage of time. It is possible that prediction conducted both before and after a course begins serves two entirely different functions. In a study, Shih and Lee [17] used the kNN algorithm to make educated guesses on the appropriateness of the study materials presented to pupils. The findings of these analyses generally demonstrated that various ML models produced excellent results and performed equally well. However, outcomes varied depending on the kind of input data used in the analysis. Furthermore, when it comes to maximizing the effectiveness of a given ML approach, the research that has been published so far has not offered a thorough examination of what constitutes the best possible input dataset.

1.2. Complexity of the Learning Process and the Role of Machine Learning

The domain of education, known for its complex and diverse characteristics, presents difficulties when just examined from an efficiency standpoint. The process of learning is not merely a straightforward progression from a state of ignorance to one of knowledge, but rather it is characterized by a diverse array of personal encounters, obstacles, and moments of enlightenment. Acknowledging this, our research aims to utilize machine learning not as a cure-all but as a tool—capable of illuminating patterns, trends, and potential areas for intervention.

It is of utmost importance to underscore that the objective of this study is not to categorize students into binary groups only based on their performance in examinations. The objective, in this case, is to utilize data in order to extract valuable insights that can contribute to the comprehensive learning experience of students. Machine learning offers a comprehensive framework for performing predictive analytics. However, it is imperative to adopt a critical perspective when interpreting its outcomes, recognizing that it mostly captures trends and patterns rather than encompassing the total of an individual's educational progress. In an effort to achieve a harmonious approach, we aim to acknowledge both the advantages and disadvantages of machine learning within the realm of educational analytics. This entails utilizing data-driven insights to improve pedagogical tactics, while also remaining sensitive to the distinct and varied experiences of each learner.

1.3. Research Objective

This study aims to examine the academic performance of students and ascertain the factors that influence their success or failure in learning. The task at hand pertains to the utilization of diverse feature engineering methodologies and machine learning models to

forecast the weighted score of pupils, taking into account multiple features such as the weight and score of assignments, module presentation, and the total recorded weight of the module. The primary objective is to offer valuable perspectives that can assist educators in recognizing students who are at risk and equipping them with the essential resources to excel in their academic endeavors.

1.4. Research Contribution

This study aims to examine and forecast student performance through various feature engineering techniques and machine learning models. Consequently, this research could make a noteworthy contribution to the field of education. The research outcomes have the potential to assist educational professionals and decision-makers in detecting students who are susceptible to suboptimal academic achievement. This could enable them to offer appropriate interventions and assistance to enhance their academic performance. Furthermore, the methodology and findings of this study may offer valuable insights and direction for future research endeavors in the realm of analyzing and predicting student performance. This study has the potential to enhance educational practices and outcomes, thereby making a valuable contribution to the field.

2. Literature Review

Predicting a student's performance is one of the most essential factors in speculating students' career prospects. Not only will the prediction assist the students, but various agencies also require efficiency in academic management such as student retention, admissions, and alumni relations, as well as facilitating more precise and relevant advertising and promotion. School-based intervention initiatives can benefit the prediction of potential kids at risks. To be effective, these types of programs need to identify the kids in need of support quickly and accurately and place them in order of priority. This section provides a comprehensive literature analysis covering the years 2010–2022, focusing on papers that show the effectiveness of ML methods in improving the academic outcomes of students at risk. Studies are also described that pertain to the dataset type, feature selection techniques, classification criteria [18], testing methodologies, and results of the suggested approaches.

2.1. Risk Prediction in Student Performance Employing Machine Learning

In order to study student behavior in VLE systems, Kuzilek [19] relied on General Unary Hypotheses Automation (GUHA) along with Markov Chain-based exploration. There were thirteen situations in the collection. This evaluation relied on a dataset that included (a) students' grades for assigned work and (b) a record of students' activities inside a virtual learning environment (VLE). The LISP-Miner program was used to carry out the analysis. After performing their research, they determined that either approach would provide useful results when mining the dataset. A graphical model built on the Markov Chain may make the information more readily apparent. The intervention strategy is bolstered by the patterns collected utilizing the techniques. Students' future academic success may be forecast using data on their past behaviors.

He [20] presented the risk detection in massive open online courses. They presented two transfer learning algorithms: LR-SEQ for sequential data and LR-SIM for simultaneous data. DisOpt 1 and DisOpt2 data were used to test and compare the effectiveness of the suggested algorithms. When compared to the original Logistic Regression (LR) model, LR-SIM had a higher AUC value in the first week, making it the clear winner when comparing the results to the LR-SEQ. This finding reflected a positive forecast during the preliminary admissions process.

Kovacic [21], utilizing ML methods, investigated the possibility of advanced prediction of student progress. The review analyzed the relationship between socio-demographic factors (such as education, employment, gender, marital status, handicap) and instructional factors (such as course program and course block) for the sake of accurate forecasting. These dataset characteristics were gathered from the Open University of New Zealand.

Feature-selection methods were developed for ML to zero in on the most important factors influencing student performance. The analysis concluded that course program, ethnicity, and course block are the most significant factors influencing students' performance.

Kotsiaritis [22] suggested the use of a method called the combinational incremental ensemble of classifiers to forecast student performance. The suggested method utilizes a consortium of three classifiers, each of which contributes to the final prediction. The ultimate forecast is determined by popular vote. In the case of data that are being created incessantly, this method is useful since each classifier may make a prediction about the result of a new sample as it is received. Voting determines which forecast is chosen in the end. Hellenic Open University provides the data used for training in this study. There are a total of 1347 instances in the writing-grades dataset, and each of those instances has four characteristics and four features for writing grades. The system was constructed through an integrated incremental ensemble of three ML models: naïve Bayes (NB), a neural network (NN), as well as WINDOW. The models are first trained with the help of the training set, and then they are put to the test with the help of the testing set. All three classifiers estimate the value when a subsequent instance of observation comes, and the most accurate one is chosen automatically.

Osmanbegovic [23] compared the accuracy of the prediction algorithms NB, NN, and DT when it came to evaluating student performance. There were two components to the data. The first chunk of information came from a poll taken in 2010–2011 by students at the University of Tuzla. All of the participants were freshmen majoring in economics. The second source of information was the enrollment database. There were 257 unique occurrences in the data, and they were all accompanied by 12 different properties. As a means of actualization, they resorted to Weka software. The performance of the classifiers was assessed by measuring their precision, speed of learning, and mistake rate. While training took less than 1 s, the NB achieved a remarkable accuracy result of 76.65% while having significant mistake rates.

While attempting to predict how students in an introductory programming course will perform, Watson [24] took into account their activity record. This analysis suggested using an indirect foundation, such as automatically assessed criteria, to track student achievement over time. They created an evaluation method for student programs called WATWIN, which gives points for different kinds of work. A student's score was based on their responsiveness to and speed in fixing programming problems. For this analysis, we utilized data from 45 students' coding activity logs throughout 14 sessions. Each student's behavior was given a WATWIN score, which was subsequently employed in linear regression analysis. In the linear regression approach, the WATWIN score produced an accuracy of 76%. Prediction accuracy depends on a well-balanced dataset. In a balanced dataset, there is an equal attribute for each of the classes used in the prediction.

Hu [25] looked at the factors that change over time and their ability to predict online students' success. They suggested setting up an electronic warning system to identify at-risk pupils in asynchronous learning environments. They suggested using time-dependent factors as a crucial criterion in evaluating LMS students' progress. This paper's primary aims were to (i) explore data mining techniques for advanced warning, (ii) determine the implications of time-dependent factors, and (iii) choose data mining techniques with greater predictive ability. Three ML classification models, CART, LR, and AdaBoost, were tested using data from 330 learners enrolled in online courses in the LMS. Each instance in the dataset included 10 attributes, and the classifiers' effectiveness was measured by their accuracy as well as their resistance to making type I and type II mistakes. With an accuracy rate of 95% or above, the CART algorithm easily beats out the competition.

Lakkaraju [26] developed an ML framework to predict which students would be most likely to drop out of school or who would take too long to get their degrees. Information on students from two different schools and two different school districts is gathered using this structure. SVM, RF, LR, Adaboost, and DT are the five ML models put to use in experiments. Precision, recall, accuracy, and area under the curve (AUC) for binary classification are used

to rate these methods. Every student is given a ranking that takes into account his/her assessed level of risk according to the model of categorization presented above. Based on the findings, Random Forest is the most effective method. The algorithms are ranked according to their accuracy and recall. Based on the suggested framework, the authors may recommend five important procedures for teachers to carry out in order to avoid the most common blunder. Some of these are (a) employing the FP-Growth algorithm to find patterns in the data, (b) employing the score's risk to rank the students, (c) adding a new field to the data and giving it a value of one (1) if the framework failed to predict it correctly and zero (0) otherwise, (d) computing the probability of error for each of the frequent patterns, and (e) sorting the patterns by error probability.

Ahmed [27] sifted through the student database at the school for information on its students. There were 1547 instances in the collection, and each had 10 characteristics. Various departments, high school diplomas, grades at midterm and on lab tests, seminar and assignment grades, seminar and assignment engagement and performance, seminar and assignment scores, homework and final grade averages, and final grade averages were all collected using the qualities chosen. Data categorization utilized DT and ID3 Decision Tree, two ML models. Then, the Weka data-mining tool was used to conduct experiments. With the use of the collected data, they determined that the midterm property should serve as the tree's starting point.

In contrast to other research, Marbouti [28] studied prediction models to detect underperforming students in a standard-based education setting. Additionally, they used the data from the first-year engineering program at a Midwestern US University in 2013 and 2014 to implement feature selection techniques that reduced the feature space. The student progress data collection included things like quiz scores, homework grades, teamwork marks, project checkpoints, quantitative modeling, and exam results. Classifiers from KNN, LR, NB, SVM, MLP and DT were among the six classifiers that were chosen in this study. Overall accuracy, pass-rate accuracy, and failure-rate accuracy were used to assess the performance of these classifiers. Features having a Pearson's correlation coefficient value greater than 0.3 were employed in the prediction procedure as part of the feature selection approach. In contrast, NB models trained with 16 characteristics performed better (88 percent accuracy).

Additionally, Iqbal [29] predicted learners' GPA utilizing CF, MF, and RBM in a comparable assessment. This analysis relied on data gathered from Information Technology University (ITU) in Lahore, Pakistan. A feedback model was presented to determine the extent to which a student had grasped the material covered in a given class. In addition, they proposed a method for tuning the Hidden Markov model (HMM) such that it may be used to foretell how well students would perform in a given class. The experimental data were divided into a 70% training set and a 30% testing set. RMSE, MSE, as well as Mean Absolute Error were used to rank the ML-based classifiers. For this dataset, RBM performed well, with mean squared errors (MSEs), Mean Absolute Errors (MAEs), and Root-Mean-Squared Errors (RMSEs) all below 0.

Almarabeh [30] looked at how well various classifiers measured up while analyzing student work. In this analysis, they looked at five different ML-based classifiers and compared their features. Some examples of classification models include NB, J48, Bayesian Networks, NN, and ID3. All of these methods were implemented in Weka and utilized in the tests. Data for this study were culled from a college database containing 225 cases, each of which had 10 characteristics. This analysis demonstrated that the Bayesian Network is the best prediction method for real-world use.

Xu [31] proposed a novel ML technique with two standout characteristics. The first was a multi-tiered framework for forecasting that takes into account pupils' dynamic performance behaviors. Multiple bases as well as ensemble classifiers composed the layered structure. The suggested method's second key aspect was its data-driven strategy for determining lecture topics. Records for 1169 UCLA undergraduates were included in

the collection, which focused on the departments of Aerospace and Mechanical Engineering. In terms of MSE, the suggested technique performed well.

Using data collected at the University of Minho and consisting of 33 characteristics, Al-shehri [32] conducted a similar review in which he compared the efficacy of SVM and KNN supervised learning classifiers. Before performing statistical analysis, the dataset was transformed from categorical to numeric forms. Questionnaires and teacher reports from two Portuguese schools formed the basis of the dataset. There were 395 unique occurrences in the original dataset, and its 33 characteristics included nominal, binary, and numeric attribute distributions of 4, 13, and 16, respectively. In the experiment, several data division sets were employed to test the methods, and the Weka program was used to do so. When utilizing a 10-fold cross-validation and partition ratio, the SVM was shown to have a good degree of accuracy.

Alowibdi [33] investigated how advanced learning analytics may be used to forecast student success. Scholars in Pakistan were taken into consideration. In this study, they compared the Bayes Network and NB against the discriminative CART, SVM, and C4.5 models. Measures of precision, recall, and F-score were used to assess the efficacy of the predictors. From 2004 to 2011, data for 3000 students were gathered; following pre-processing and redundancy reduction, this number was reduced to 776 learners. Only 690 of the 776 students were able to completely finish their degree programs. Expenditure, income, student demographics, and asset data were separated out of the total of 33 factors. Spending on natural gas, electricity, self-employment, and location emerged as the most significant factors predicting student success in the research. With an F1 score of 0.867, the SVM classifier easily beats out the competition.

In order to determine what factors into the PISA 2005 result, Msaci [34] used ML methodologies. In this work, the author examined PISA 2005 data from a wide range of countries, including the United States, the United Kingdom, Australia, Spain, Canada, Italy, Japan, and France. The purpose of this research was to identify the factors at play in both the students' and the schools' environments that have a bearing on students' performance, with the suggested methods operating in two stages. First, they used a multilevel regression tree that nested individuals inside their respective schools to identify factors at the student level that are predictive of academic performance. After that, they estimated schools' added value to better understand what factors are associated with student success via the use of regression trees and boosting methods. The PISA 2015 dataset was utilized, which included data from nine countries and included a total of 19 characteristics (eighteen at the school level and one at the student level).

Aggarwal [35] used eight distinct ML algorithms to compare academic aspects and to evaluate the importance of nonacademic variables, such as demographic information. They used data collected from an Indian technical college, which included demographics, course information, and student activities for a total of 6807. To remedy the asymmetry of the data, the team used synthetic minority oversampling techniques. They asserted a 93.2% F1 score using the DT, 90.3% using the Lr model, 91.5% using the MLP model, 92.4% using the SVM model, 92.4% using the AdaBoost model, 91.8% using the Bagging model, 93.8% using the RF model, and 92.3% using the Voting model. The authors also argued that students' success in school is affected not only by their academic characteristics, but also by their socioeconomic background, and they proposed combining non-academic factors with academic ones to predict students' outcomes.

Using AutoML, Zeineddine [36] improved the accuracy of student performance prediction by capitalizing on characteristics collected in the lead-up to the commencement of the new curriculum. With AutoML, they were able to reduce the false prediction rate while maintaining an accuracy of 75.9% and a Kappa value of 0.5. They concluded that AutoML should be used by researchers in this area to find the best possible learner performance prediction model, particularly when starting with pre-start data. To support students in imminent need of assistance, they recommended using pre-admission data to initiate intervention and consultation sessions prior to beginning academic progress. Because of the

uneven distribution of the available data, they used the SMOTE pre-processing approach and the automatically created ensemble methods to make accurate predictions about which students will fail. The authors recognized SMOTE's overgeneralization weakness and suggested ways to mitigate the unbalanced data issue without resorting to a less than ideal solution: a more balanced dataset.

Using past data, Bueno-Fernández [37] advocated using ML techniques to predict learners' ultimate grades. They used information gathered from the universities of Ecuador's computer engineering departments. The collection of a large amount of high-quality data was the primary objective of this study. Their strategy produced a plethora of information that, with proper processing, could be repurposed into several beneficial tools for the field of education. This study offered a unique method for pre-processing and clustering students with similar patterns. After that, they used a wide variety of supervised learning techniques to zero in on the students whose patterns were most similar and to forecast their final grades. The findings from ML approaches were then compared to the state of the art. They asserted a 91.5% accuracy using ensemble methodologies, demonstrating the efficacy of ML approaches to predicting students' performance.

Many academics, Reddy and Rohith noted, had used the sophisticated ML algorithms to accurately forecast student performance; nevertheless, they had failed to provide any helpful suggestions for students who were struggling. In order to overcome this barrier, they set out to uncover what individual factors can predict a student's poor performance in a classroom setting. With the help of DT, SVM, GB, and RF, data from the University of Minnesota were analyzed. They asserted a higher rate of accuracy (over 75%) in determining which students would fail this term based on a set of characteristics that are broad enough to apply to all of them.

In another study undertaken by Kuzilek [19] at the United Kingdom's The Open University, the researcher devised a method based on three prediction algorithms to identify at-risk pupils. By analyzing two datasets, each of the three ML models K-NN, NB, and CART generated a predicted score. One was a demographic dataset culled from the school's main database, and the other was a log dataset containing information about how students and teachers interact with one another in the VLE. Each student's final grade was the total of their prediction scores from all of the algorithms used. If the total score was more than 2, it was decided that the pupil was at danger and the necessary steps were taken. If a student's cumulative score was below 3, however, they were not at danger and did not need assistance. The suggested system's efficacy may be measured, in part, by its precision and recall scores.

In recent years, the EDM research community has focused a lot of its emphasis on e-learning systems. Hussain [38] looked into machine learning techniques to foresee problems with the e-learning platform Digital Electronics Education and Design Suits (DEEDS) that students would run into. EDM strategies take into account the student's engagement with the system to reveal relevant patterns that may inform and enhance instructional decisions. Here, information utilization was carried out from one hundred incoming BSc majors at the University of Genoa. Session logs generated by students' use of the DEEDS tool may be accessed by the general public via the UCI ML repository. The average time, the total amount of activities, the average idle time, the average number of important storks, and the total number of connected activities were the five variables chosen to predict student achievement. The five ML techniques covered in this review were artificial neural networks, support vector machines, Decision Trees, and deep learning. The classifiers' efficiency was measured using the Root-Mean-Squared Error metric (RMSE), the ROC curve, and the Cohen's Kappa coefficient. Additional measures of effectiveness included accuracy, precision, recall, and F-score. Findings in terms of RMSE and evaluation metrics were equal for ANN and SVM. The authors argued for the use of support vector machines (SVMs) and artificial neural networks (ANNs) in predicting student performance, and they presented a modified DEEDS system that included ANN and SVM algorithms.

2.2. Students Dropout

Alhusban [39] used an ML study to track and improve retention rates among college learners. Al-Al Bayt University students' gender, enrollment type, entrance marks, place of birth, nationality, courses, and marital status studied during elementary and secondary school were among the variables assessed. They used Hadoop, an open-source platform based on ML, since there are so many features incorporated that the sample dataset is large. The results of the admissions exam were discovered to have a substantial impact on the chosen field of study. Furthermore, they argued that particular sexes tend to dominate certain spheres, such as the medical industry, where many more females than boys choose to specialize. They further argued that pupils' socioeconomic backgrounds had an impact on their academic success. As a conclusion, they said that single learners outperformed their married or coupled counterparts.

The data for the Yukselturk study [40], which were acquired via online surveys, were reviewed by Yukselturk in order to evaluate the data mining approaches for dropout prediction. The online questionnaire consisted of 10 sections, which were titled as follows: age, profession, gender, self-efficacy, education level, prior knowledge, coverage, previous online experience, and locus of control. There were a total of 189 students that participated in the event. The investigation made use of four different ML models and a strategy that was based on the genetic algorithm to identify features. According to the findings, 3NN was the most effective classifier, with an accuracy rate of 87%.

The authors of [41] provided some insight into the role of temporal characteristics in the prediction of student dropout. Using information gained from student quiz scores and data collected from networking forums through the Canvas API, the temporal features were able to capture the changing patterns of student performance over time. The following are some of the attributes that were retrieved from the data: active days, number of module views, number of forum views, dropout week, number of quiz views, number of discussion posts, and social network degree. Both the BN and the DT were used as classification strategies throughout the classification process.

Liang and Zhen [42] examined information about students' classroom participation to calculate the likelihood of student withdrawal in the subsequent days. Gathering information from the XuetangX forum, pre-processing that information, extracting and selecting features, and using machine learning techniques were all parts of the proposed framework. There were 39 Open-Edx-based courses included in the XuetangX online learning dataset. Throughout a 40-day period, pupils recorded their actions and attitudes. In order to train ML algorithms, the raw log data must first be processed. There were a total of 121 characteristics retrieved, and they were split evenly across three groups: users, courses, and enrollments. The dataset was then split into a training set with 120,054 cases and a testing set with 80,360 instances. When the average area under the curve (AUC) for a certain classifier was large for a GBT, SVM and RF classifiers were also utilized. There were two types of information used to forecast pupils' future performance: (a) static information and (b) information that changes over time. Thaker [43] presented a framework for adaptable textbooks based on a dynamic model of student knowledge. The dynamic learner's performance data reportedly included learner success and failure diaries compiled as they engage with the LMS. As the features of the dataset vary over time, learner interaction sessions with the e-learning platform are an illustration of dynamic data. Alternatively, dynamic learner performance data are continuously updated, whereas static data are obtained just once. Enrollment and demographic information about students is one such example. As a solution, the suggested framework takes into account students' reading habits and quiz scores to estimate where their knowledge is at the moment. The framework has two more complex forms of the standard Behavioral Model (BM): the Behavior-Performance Model (BPM) and the Individualized Behavior-Performance Model (IBPM). The suggested models were implemented using the Feature Aware Student Knowledge Tracing (FAST) tool. The suggested method outperformed the standard Behavior Model in terms of RMSE and ACU.

Carlos [44] introduced an ML classification model for predicting learner performance; this model incorporates a data gathering strategy to glean information about student learning and behavior from instructional settings. Students' performance was utilized to divide them into three groups using the support vector machine (SVM) algorithm: high, medium, and low. The author gathered information from 336 pupils across 61 dimensions. The first experiment utilized just behavioral characteristics for classification, the second experiment only used learning features, the third experiment mixed learning and behavioral features for classification, and the fourth experiment used only chosen features for predicting student success. Predictions of student performance over a ten-week period were based on the general eight behavioral variables and 53 learning factors included in the dataset. The findings demonstrated that the classifier's performance improved each week as more data were collected. In addition, by week 10, a combined effort of behavioral and learning variables yielded a high classification result of 74.10%.

3. Methodology

The complete methodology for student learning analytics in academia (SLAIA) involves several steps, which are depicted in Figure 1. The first step is to define the collection of the OULA (open university learning analytics) dataset and preprocess it. Data cleaning, handling of missing values, noise reduction, and feature encoding are performed here to make the OULA dataset clean and ready for the predictive analysis. In this study, the regression and classification are performed to forecast the student performance and build a robust predictive model for SLAIA. Feature engineering is also performed here to uncover the most influential features from the OULA dataset. Two feature engineering techniques are used for regression (LASSO, Boruta) and two techniques are used for classification: Random Forest Importance (RFI) and Recursive Feature Elimination (RFE). The main objective is building the predictive model based on the best subset of features for SLAIA. Furthermore, the four machine learning (ML) techniques are trained and tested for classification and regression. The ML model names for classification are ensemble model (EM), XGBoost (XBG), neural networks (NNs), and Decision Tree (DT), whereas the regression analysis model are linear regression (LR), support vector regressor (SVR), Random Forest Regressor (RFR), and Gradient Boosted Regressor (GBR). The evaluation of both ML analyses is also performed using the evaluation metrics to find the best ML techniques for classification and regression analysis. This proposed framework will find the best ML models for each analysis and find the best influential features for the SLAIA to assist the learners and academic professionals. In general, the methodology utilized for student learning analytics entails a methodical approach to gathering, scrutinizing, and comprehending data with the aim of enhancing educational achievements. Through the utilization of insights based on data, educators and academic institutions can make well-informed decisions aimed at promoting and enhancing student achievement.

3.1. Dataset Description

The open-source open university learning analytics (OULA) dataset from the Kaggle platform is considered for this research. Behavioral and performance are two aspects of the framework's components offered by the OULA dataset. It includes details on 32,593 learners, 22 courses of study, the assessment scores, and records of their VLE (virtual learning environment) conversations, which include each day's summary of student clicks. Learner course engagement behaviors, efficiency, and comfort are considered in the OULA dataset. Analytics through LMS along with additional education and technological resources are used to collect behavioral data. Student assessments are used to compile performance statistics and surveys are used to acquire satisfaction information. The OULA dataset schema is listed in Figure 2.

There are a total of "7" categories, and their names and the description of each feature from the OULA dataset are listed in Table 1.

The variables chosen for analysis in the OULA dataset were determined to be both accessible and pertinent to the objectives of this study. The variables used for this study were influenced by both the availability of data and the literature and empirical findings that indicate their importance in predicting student success.

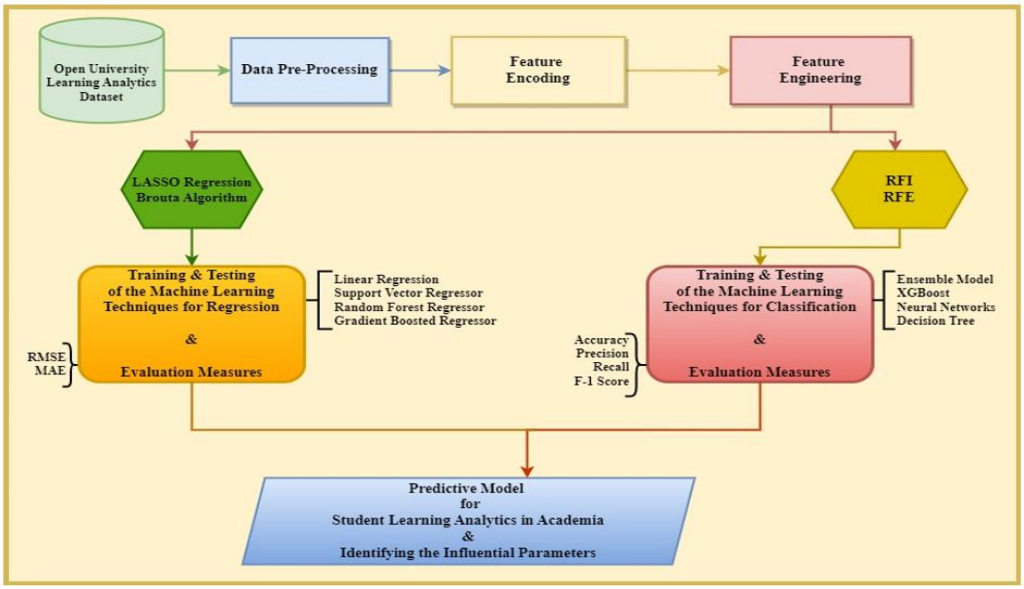


Figure 1. Methodology framework for the SLAIA.

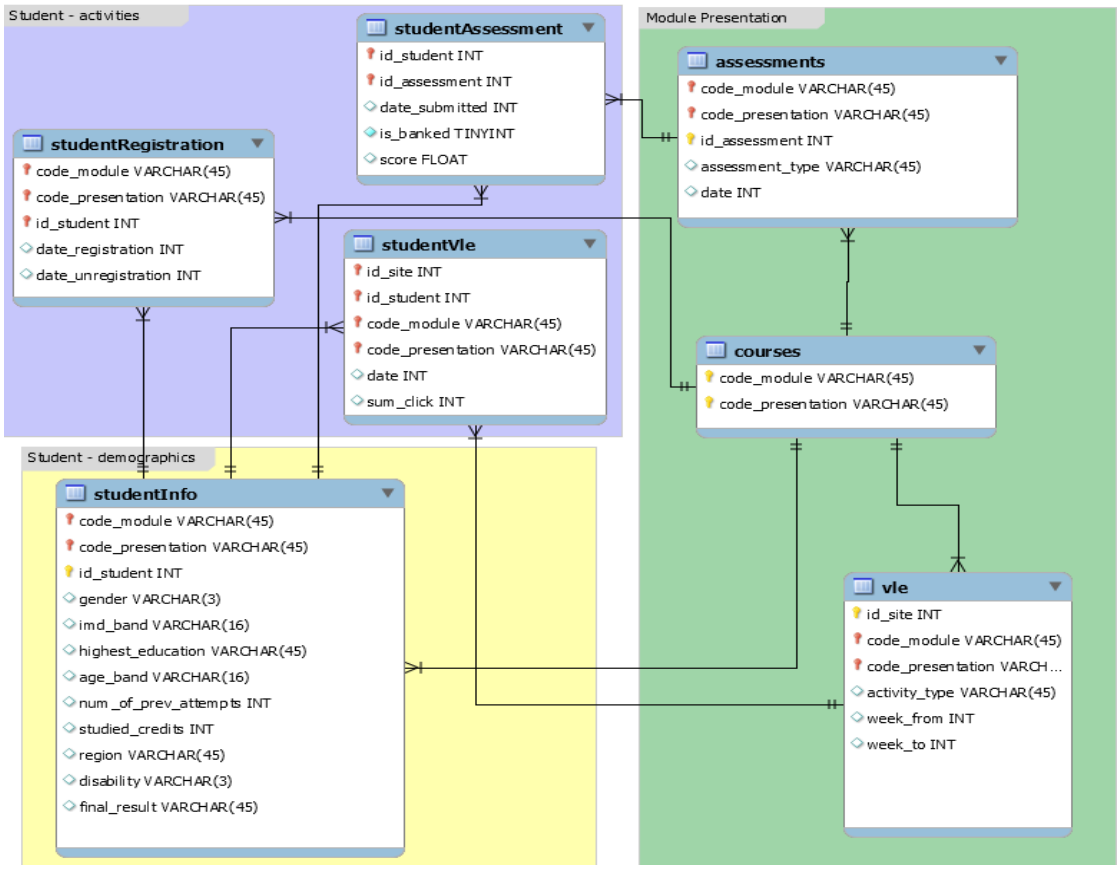


Figure 2. OULA database schema (https://analyse.kmi.open.ac.uk/open_dataset#about (accessed on 9 March 2023)).

Table 1. Description details of the OULA dataset.

S. No.	Feature Category	Feature Name	Description
1	Courses	code module	Distinctive ID that is assigned to a module
		code presentation	Refers to the designated identifier for a given presentation
		length	Represents the duration of the module presentation
2	Assessments	code module	An identifying code for the part of the module of which the assessment is associated
		code presentation	Denotes the identifier code of the document being presented to which the assessment pertains
		ID assessment	The unique identifier which is assigned to a particular assessment
		Assessment type	The specific type of evaluation being conducted
		Date	Depicts the final delivery deadline of the assessment
		Weight	The percentage of importance assigned to an assessment
3	VLE	ID site	A unique identifier that is assigned to a particular piece of material
		code module	Serves as an identifying code for a particular academic module
		code presentation	The unique identifier that is assigned to a particular presentation
		Activity type	Refers to the specific role that is linked to the material within the module
		Week from	The designated week for which the instructional material is intended to be utilized
		Week to	The duration for which the instructional material is intended to be utilized
4	Student Information	code module	A unique identifier assigned to a specific academic course in which a learner is enrolled
		code presentation	Refers to the unique identifier assigned to the presentation for which the learner is enrolled
		Id Student	A distinct identifier for each individual student
		gender	Refers to the sex of the student
		region	The specific geographic location in which the student was residing during the module presentation
		highest education	The highest level of education attained by the student at the time of enrollment in the module presentation
		imd band	The Index of Multiple Deprivation bands is associated with the geographic location of the student
		age band	Represents the range of ages to which the student belongs
		No of pre-attempts	The frequency of attempts made by the student for the current module
		studied credits	The aggregate number of credits assigned to the modules that a student is presently enrolled in
		disability	The declaration of a disability by the student
		final result	The student's ultimate outcome in the module presentation
5	Student Registration	code module	A unique identifier assigned to a specific module
		code presentation	Specific code assigned to a particular presentation
		Id Student	A distinct identifier for each individual student.
		date registration	The date on which a student enrolled for a module presentation
		date unregistration	The duration, in days, between a student's enrollment and unregistration from a module

Table 1. *Cont.*

S. No.	Feature Category	Feature Name	Description
6	Student Assessment	id assessment	The unique identifier assigned to an assessment
		id student	A distinct and exclusive numerical representation for each individual student
		date submitted	The date on which a student submitted their work
		is banked	A status indicator denotes the transfer of an assessment outcome from a prior presentation
		score	The numerical value represents the performance of the student in the given assessment
7	Student VLE	code module	A unique identification code is assigned to a module
		code presentation	The identification code that is assigned to the module presentation
		id student	A distinct identification number that is assigned to each student
		id site	A unique numerical identifier that is assigned to the VLE material
		date	Student's engagement date with the material
		sum click	The quantification of a student's engagement with the learning material within a given day

The selection of variables is undoubtedly a critical factor in determining the outcomes and implications of any study that relies on data analysis. The variables played a fundamental role in the construction of the predictive models in our study. The utilization of distinct variables may lead to variations in the performance of the models. In order to enhance the reliability and pertinence of our findings, we endeavored to choose variables that have been previously associated with student performance.

3.2. Data Preprocessing

Data preprocessing played a crucial role in guaranteeing the validity and dependability of the outcomes in our investigation. The OULA dataset was subjected to a rigorous cleaning process, wherein missing data were carefully handled to eliminate any null values. In order to address the potential bias caused by anomalies, noise reduction techniques, such as the Interquartile Range (IQR), were implemented. It is imperative to acknowledge that the handling of missing data is not only crucial for maintaining the integrity of the model but also for ensuring the generalizability of this study across other institutions. The absence of a standardized approach to data treatment poses a potential challenge when making comparisons between institutions and colleges, as it may lead to distorted or less dependable results. The approach employed places a strong emphasis on maintaining data integrity, hence ensuring the production of robust results. This characteristic allows for the adaptability of the methodology in many educational environments.

3.3. Performance Evaluators for Student Learning Analytics in Academia (SLAIA)

Two performance evaluators for classification and regression analysis in SLAIA prediction are formulated. The final weighted score is computed for the regression analysis and the final results are determined for the classification problem.

3.3.1. Final Results

The final outcome of a student's performance in a module presentation is called their final results. If the module has distinction, it is converted into "pass", whereas the withdrawal from assessment results is considered as "fail". It is a binary classification problem, and it has two levels that belong to pass and fail of the assessment of the various students.

3.3.2. Final Weighted Score

The final weighted score is used for the regression problem and its formulation is given using the description below with equations. The process of calculating the weighted score for a module in a course is as follows: Each assignment is assigned a weight (w) and a score (s), and the weighted score for each assignment is calculated as the product of w and s. The sum of all weighted scores for a module is divided by the total recorded weight of the module (tw_m) to obtain the average weighted score for that module. The module presentation (p) is also added to the sum, and the entire expression is divided by the total weight of the module and presentation ($m + p$) to obtain the final weighted score. This equation allows educators to assign appropriate weights to assignments and calculate an accurate assessment of a student's performance in a course.

Then, the mathematical equation for the given process would be

$$\text{weighted}_{\text{score}} = \text{w} \times \text{s} \tag{1}$$

$$sum\_weighted\_score = \frac{\sum(Weighted\_Score)}{m+p} \tag{2}$$

$$\text{module\_total\_recorded\_weight} = \text{tw\_m} = \frac{\sum(w)}{m+p} \tag{3}$$

$$final\_weighted\_score = \frac{\sum(\frac{\sum(w \times s)}{\sum(w)})}{m+p} \tag{4}$$

4. Feature Engineering Methods

4.1. For Regression

Lasso regression, which is also referred to as L1 regularization, is a linear regression methodology that incorporates a penalty to the absolute values of the coefficients [45]. This penalty results in the reduction in some coefficients to zero. Lasso regression possesses a notable benefit in its capacity to conduct automatic feature selection. This is achieved through the process of coefficient shrinkage, whereby certain features are reduced to zero while preserving those that are deemed significant. The technique aids in mitigating overfitting by removing extraneous features, rendering it especially advantageous in datasets with a high number of dimensions and numerous features.

The Boruta algorithm is a technique for selecting features that are applicable to datasets with a large number of dimensions, and is capable of identifying variables that are pertinent to the analysis [46]. The method is a wrapper technique that leverages a foundational machine learning algorithm. Its functionality involves evaluating the significance of individual features in comparison to a set of shadow features that are generated randomly. The algorithm has been developed with the purpose of identifying pertinent variables that exhibit a low correlation with other characteristics. Additionally, the algorithm aims to prevent the occurrence of false positives by refraining from selecting irrelevant variables that exhibit a high correlation with other features. The hyperparameters for each model used in this study through assistance of the SKLearn framework are given below in Table 2.

Table 2. Regression model hyperparameter selection.

S. No.	Model Name	Hyperparameter Details
1	Linear Regression	fit_intercept = True, normalize = False, n_jobs = None, Positive = False
2	Support Vector Regressor	C = 1.0, kernel = rbf, degree = 3, gamma = scale, coef0 = 0.0
3	Random Forest Regressor	n_estimators = 100, criterion = "mse", max_depth = None
4	Gradient Boosted	Loss = "ls", n_estimators = 100, max_depth = 3, learning_rate = 0.1

4.2. For Classification

The Random Forest Importance (RFI) technique is a feature selection approach that leverages random forests to ascertain the significance of features for classification assignments [47]. The RFI technique computes a numerical value for each feature by measuring the reduction in the average impurity of the Random Forest model upon the removal of that feature. The features that obtain the highest scores are deemed to be of utmost significance for the purpose of classification.

The Recursive Feature Elimination (RFE) method is a machine learning technique employed to select the most significant features from a provided dataset [48]. The methodology involves iteratively eliminating features of lower significance from the dataset until the desired quantity of features is attained. This approach is frequently employed in conjunction with other machine learning techniques to enhance their efficacy. The RFE methodology operates by allotting weights to individual features present in the dataset, taking into account their significance in forecasting the output variable.

4.3. Machine Learning Model and Evaluation Measures

The present study employed classification and regression models to investigate the field of student learning analytics in academia (SLAIA). In order to perform classification, a range of machine learning models were utilized, including ensemble learning, Multi-layer Perceptron, XGBoost, and Decision Tree models. In the context of regression analysis, a set of models including linear regression, support vector regression, Random Forest Regressor, and Gradient Boosted Regressor were employed. In order to assess the efficacy of our classification models, we employed various metrics including accuracy, precision, recall, F−1 score, and confusion matrices. In the context of regression analysis, the Root-Mean-Squared Error (RMSE) and the Mean Absolute Error (MAE) are two widely used evaluation metrics. In summary, the utilization of classification and regression models facilitated the acquisition of valuable knowledge pertaining to student achievement and the determination of influential features for the SLAIA. The selection of hyperparameters for each classification model is shown below in Table 3.

Table 3. Classification model hyperparameter selection.

S. No.	Model Name	Hyperparameter Details
1	Ensemble Model: (a) KNN (b) Random Forest (c) Logistic Regression	(a) Weights = uniform, leaf_size = 30, metric = minkowski (b) criterion = entropy, max_depth = 8, max_features = auto, n_estimators = 500 (c) C = 100, penalty = l2, solver = newton-cg
2	XGBoost	max_depth = 4, alpha = 2, n_estimators = 50, objective = binary = logistic
3	Neural Networks	hidden_layer_sizes = 100, activation = relu, alpha = 0.2, max_iter = 6, learning_rate = constant
4	Decision Tree	Criterion = entropy, max_depth = 7, splitter = best, min_samples_split = 2

5. Experiments and Analysis

In this part of the study, student learning analytics in academia is evaluated based on several experiments. This portion of the research exclusively deals with different types of tests that are conducted on the students' academic learning dataset using the supervised learning approach. Feature engineering techniques are applied for both regression and classification problems. After feature engineering on the regression dataset, four different machine learning models are carried out that include linear regression (LR), support vector regressor (SVR), Random Forest Regressor (RFR), and Gradient Boost (GB). However, for evaluating the classification problem data, ensemble model (EM), XGBoost (XGB), neural network (NN), and Decision Tree (DT) models are applied for evaluating the student's academic learning analytics.

5.1. Feature Engineering for Regression

5.1.1. Lasso Regression

Using Lasso regression, we determine which characteristics contributed most to the prediction of the target class, "weighted_score", in a regression issue using this class as the target. The characteristics "code_module", "imd_band", and "total_click" have coefficient scores of 3.43, 0.31, and 9.95, respectively, making them the best predictors of the target variable. This indicates that these characteristics significantly affect the anticipated value of "weighted_score". The method's results show the success of Lasso regression in feature selection and finding the most important characteristics for prediction.

5.1.2. Boruta Feature Selection

In order to rank the most significant features in predicting "weighted_score", we use the Boruta feature selection method for the same regression issue. According to the data, the characteristics with the most importance are code_module, studied_credits, total_click, late_rate, and fail_rate, in that order. The fact that "code_module" and "total_click" are also shown to be top indicators in the Lasso regression study bolsters their significance in predicting "weighted_score". It is possible that "studied_credits", "late_rate", and "fail_rate" are also significant in predicting "weighted_score", since they are among the top five rated qualities. Even when the number of characteristics is substantial, the outcomes of Boruta feature selection show its efficacy in selecting the most significant ones.

Regarding feature engineering for regression issues, the usage of Lasso regression and Boruta feature selection has been shown to be beneficial. These results have real-world implications for boosting the precision of regression models in fields. Our research indicates that by using these two methods, a deeper knowledge of the data and the most crucial elements for prediction may be attained in the following sub-sections of the analysis part.

5.2. Feature Engineering for Classification

Random Forest Importance (RFI)

The RFI technique is utilized to determine the crucial features in a classification problem that predicts the "final_result" target class. The RFI analysis reveals that "total_click" is the most significant feature, with an importance score of 0.44. This indicates that the number of clicks a student makes in a course is the most influential predictor of their final result. Additionally, the RFI analysis identifies other important features, including "date_registration", "region", and "imd_band", with importance scores of 0.13, 0.08, and 0.07, respectively. These results offer valuable insights into the key factors that contribute to predicting a student's ultimate outcome in a course. The results and ranking as per the score as shown below in Figure 3.

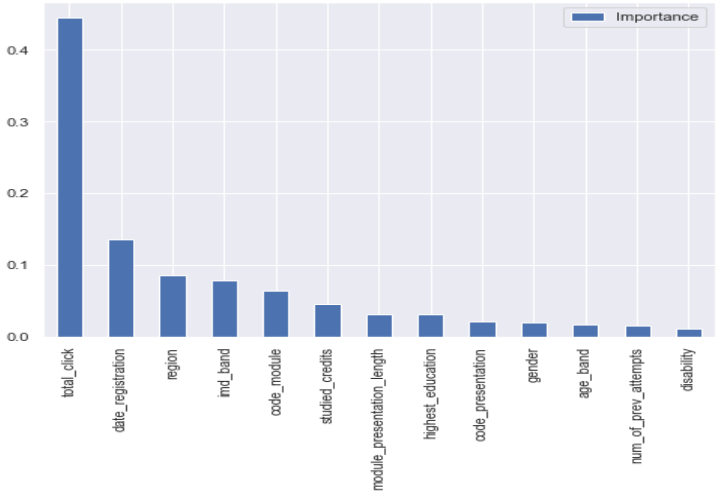


Figure 3. Feature importance illustration using RFI.

5.3. Recursive Feature Elimination (RFE)

This study also employed the RFE technique to further refine the selection of features for predicting the "final_result" target class. The RFE analysis ranked the importance of each feature based on its ability to improve the accuracy of the prediction model. The top-ranked features identified by the RFE analysis included "code_module", "date_registration", "module_presentation_length", "region", "imd_band", "studied_credits", and "total_click", all of which had a ranking score of 1. This indicates that these features were the most critical in predicting a student's final result in a course. The RFE analysis further strengthened the importance of "total_click" as the most important feature in predicting the "final_result" target class, as it was also identified as the most significant feature in the RFI analysis.

In general, the outcomes of the RFI and RFE analyses offer significant perspectives on the characteristics that hold the most significance in forecasting a student's ultimate outcome in a given course. The aforementioned discoveries can facilitate the creation of precise prognostic models and enable educators to promptly recognize students at risk of academic failure, thereby furnishing them with the requisite assistance to excel in their coursework. The amalgamation of RFI and RFE methodologies is a proficient strategy for selecting features in classification predicaments.

5.4. Student Learning Analysis Using Regression

In order to assess the influence of various features on the prediction of the "weighted_score" target variable in a regression task, this research employs four distinct machine learning models: LR, SVR, RFR, and GB. The Lasso and Boruta feature selection techniques are employed to identify the optimal features for each model.

5.4.1. Lasso Regression

For Lasso regression, the selected features are "code_module", "imd_band", and "total_click". The feature coefficients are 3.43, 0.31, and 9.95, respectively. The findings indicate that the GB model exhibits superior performance compared to the other models, as evidenced by its lower MAE and RMSE values of 18.92 and 24.29, respectively. The RFR exhibits strong performance, as evidenced by an MAE value of 19.10 and an RMSE value of 23.95. The SVR and LR models exhibit higher values of MAE and RMSE compared to the other models, which suggests a relatively inferior level of predictive accuracy. Figure 4 depicts the results of the machine learning models' outcome achieved by employing the features extracted using the Lasso regression technique.

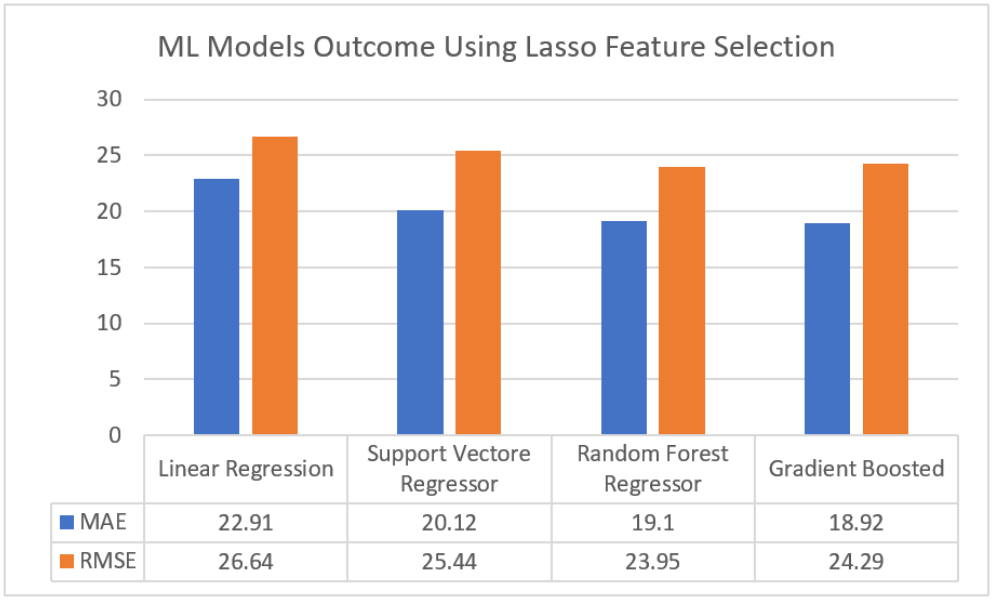


Figure 4. ML models' outcome using Lasso feature selection.

5.4.2. Boruta Feature Selection

The Boruta feature selection method yielded a set of selected features, namely "code_module", "studied_credits", "total_click", "late_rate", and "fail_rate", all of which were assigned a ranking of 1. The Boruta feature selection technique was employed to identify the optimal features for each machine learning model, namely LR, SVR, RFR, and GB. The features that were chosen for analysis included "code_module", "studied_credits", "total_click", "late_rate", and "fail_rate". Each model's performance was assessed using two evaluation metrics: MAE and RMSE. Figure 5 illustrates the results of the four machine learning models when utilizing the Boruta feature selection technique.

The findings indicate that the Gradient Boosted model exhibits higher accuracy than the other models, as evidenced by its lower MAE and RMSE values of 12.93 and 18.28, respectively. The RFR model exhibits satisfactory performance, as indicated by its MAE value of 13.52 and RMSE value of 18.46. The SVR and LR models exhibit higher values of MAE and RMSE than the remaining models, suggesting a lower level of predictive accuracy.

The findings suggest that the utilization of Lasso and Boruta feature selection methods in conjunction with GB and RFR models can proficiently predict the "weighted_score" dependent variable. This can aid educators in recognizing students who are at risk and equip them with the essential assistance to excel in their academic endeavors. The Lasso regression model is employed with the RFR and GB techniques, resulting in an MAE of 18.92 and 19.1 and an RMSE of 24.29 and 23.95, respectively. Conversely, utilizing the Boruta feature selection method in conjunction with GB and RFR results in Mean Absolute Errors of 12.93 and 13.52 and RMSEs of 18.28 and 18.46, respectively.

In our study, both MAE and RMSE values are quite similar for various models, which might make it challenging to distinguish the "best" model based solely on these metrics. Since both MAE and RMSE are measures of error, smaller values are generally better. However, the choice between them can also depend on the particular distribution of errors in your data and what you consider to be more important: error magnitude or error variance. Given that the Gradient Boosted model with the Boruta feature selection technique provides the lowest MAE and RMSE in regression tasks, and the XGBoost model yields the highest accuracy in classification tasks, these models can be considered the best performing based on these metrics.

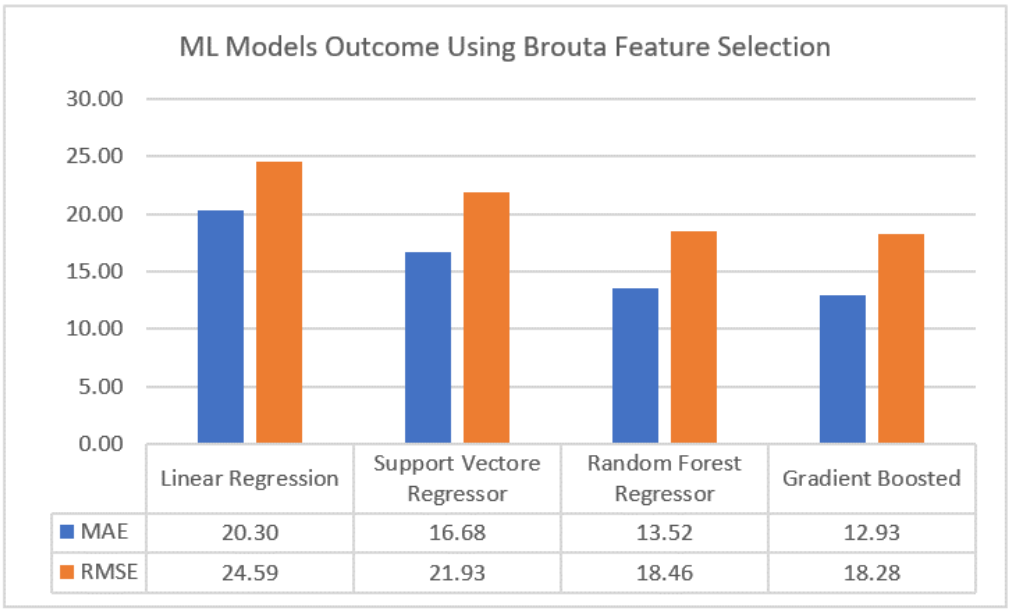


Figure 5. ML models' outcome using Boruta feature selection.

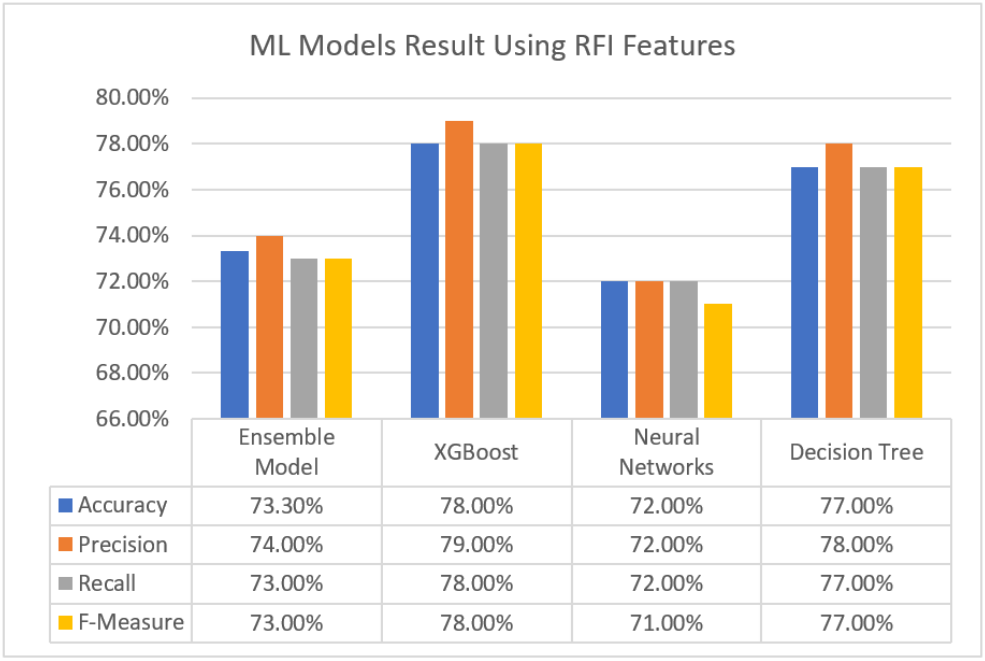
5.5. Student Learning Analysis Using Classification

Utilizing RFI and RFE to select features, this study employed four distinct classification models to forecast the target class "final_result". The utilized models included ensemble, XGB, NN, and DT.

5.5.1. Random Forest Importance (RFI)

The RFI findings indicate that the XGBoost algorithm exhibited the highest performance, with an accuracy of 78.00%, precision of 79.00%, recall of 78.00%, and F-measure of 78.00%. The DT model demonstrated favorable performance, achieving an accuracy of 77.00%, precision of 78.00%, recall of 77.00%, and an F-measure of 77.00%. The ensemble model and NN exhibited a marginally reduced level of efficacy, with an accuracy rate of 73.30% and 72.00%, correspondingly. The results are briefly presented below in Figure 6.

The findings indicate that using RFI for feature selection and implementing XGB and DT models are a viable approach for accurately predicting the target class within the context of student learning analytics classification problems. This study's results can assist educators in identifying students who may benefit from supplementary support and interventions to enhance their academic achievements.



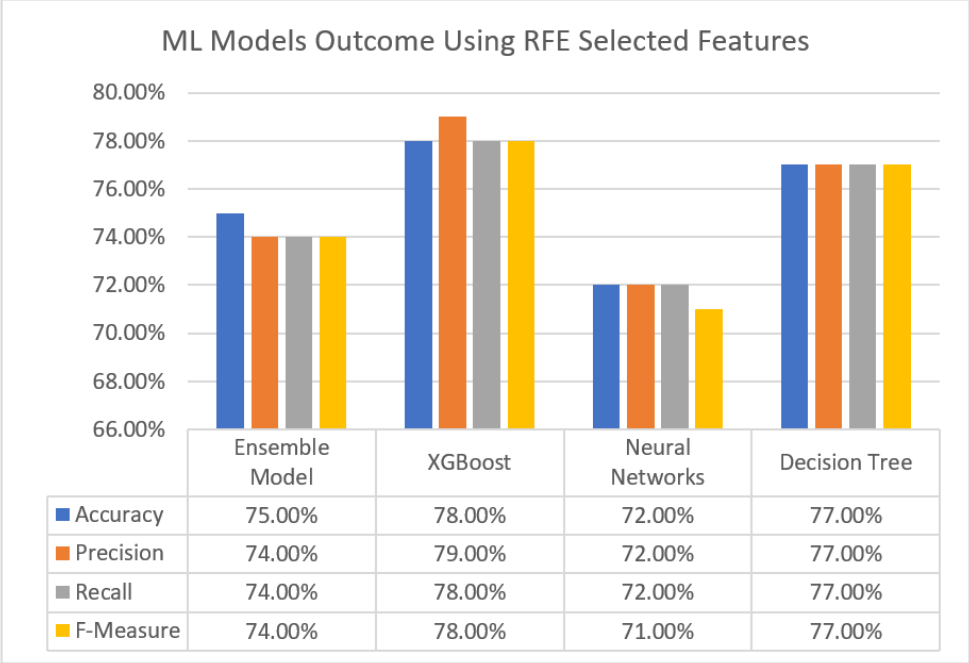
	Ensemble Model	XGBoost	Neural Networks	Decision Tree
Accuracy	73.30%	78.00%	72.00%	77.00%
Precision	74.00%	79.00%	72.00%	78.00%
Recall	73.00%	78.00%	72.00%	77.00%
F-Measure	73.00%	78.00%	71.00%	77.00%

Figure 6. ML models' outcome using RFI feature selection.

5.5.2. Recursive Feature Elimination (RFE)

The findings derived from the RFE approach indicate that the XGB and DT models exhibited efficacy in forecasting the target "learning outcome" category. Both models attained an accuracy rate of 78% and 77%, correspondingly. The ensemble model exhibited a satisfactory performance, achieving an accuracy rate of 75%. Nevertheless, the performance of the NN model was comparatively inferior to the other models, exhibiting a precision rate of merely 72%. The outcome of the ML models is depicted below in Figure 7.

In summary, the findings indicate that the utilization of the RFE feature selection method in conjunction with XGB and DT models yield favorable outcomes in forecasting the "learning outcome" classification category. The aforementioned discoveries may aid educators in detecting students who are at a higher risk of academic failure and equip them with the essential assistance to excel in their academic endeavors.



	Ensemble Model	XGBoost	Neural Networks	Decision Tree
Accuracy	75.00%	78.00%	72.00%	77.00%
Precision	74.00%	79.00%	72.00%	77.00%
Recall	74.00%	78.00%	72.00%	77.00%
F-Measure	74.00%	78.00%	71.00%	77.00%

Figure 7. ML models' outcome using RFE selected features.

In the face of closely aligned performance metrics—MAE and RMSE for regression tasks, and Accuracy, Precision, and Recall for classification tasks—determining the superior model demands a multifaceted approach. Statistical tests, such as paired t-tests, can offer insights into the statistical significance of the seemingly marginal differences in MAE or RMSE, thereby providing a scientifically grounded basis for model selection. Additionally, the quantification of relative errors and confidence intervals for these metrics can further delineate the models' performance nuances. Domain-specific criteria also come into play; in an educational context, the weightage given to different types of errors—be it false positives or false negatives—can tip the scale in favor of a particular model. Lastly, computational efficiency remains a pragmatic tiebreaker when performance metrics are too similar to definitively favor one model. Incorporating these diverse criteria provides a holistic, scientifically rigorous framework for model selection, elevating the analysis beyond the limitations of raw metrics alone.

In our comprehensive assessment of predicting student performance, the Boruta feature selection technique combined with the Gradient Boosted Regression model emerged as the standout, achieving an impressive MAE of 12.93 and RMSE of 18.28. On the classification front, irrespective of the feature selection technique employed, the XGBoost model consistently outperformed its counterparts, registering an exemplary accuracy rate of 78%. These findings underscore the potency of a meticulous feature selection process coupled with robust machine learning models, illuminating a clear pathway for optimal predictive analytics in educational contexts.

6. Discussion

The utilization of Gradient Boosting in this research has notable benefits, including its resilience to outliers, ability to handle diverse data formats, and provision of valuable insights into the value of features. These traits are particularly well suited to the characteristics of academic performance data. Nevertheless, it is crucial to acknowledge the possible drawbacks associated with this approach. One potential limitation is the susceptibility to overfitting if the method is not adequately fine-tuned. Additionally, the computational requirements may increase significantly, particularly when dealing with a large number of trees. Furthermore, the success of the method might be very dependent on the individual

hyperparameters that are selected. These factors are essential for analyzing the effectiveness and suitability of the strategy in forecasting student outcomes.

Risks and Considerations in Applying Machine Learning to Educational Analytics

The utilization of machine learning presents a potentially fruitful route for comprehending and forecasting student performance. However, it is imperative to adopt a nuanced perspective when interpreting the outcomes it produces. An inherent risk associated with a heavy reliance on prediction models is the possibility of oversimplifying complex phenomena. By placing excessive emphasis on mean values and prevalent trends, there is a risk of unintentionally disregarding exceptional cases or aberrations. These differences, rather than being simply oddities, may potentially signify distinct learning paths or particular obstacles encountered by specific student populations. Moreover, an excessive focus on predictive analytics may accidentally result in educators giving priority to conforming to the model, which could potentially marginalize the wider educational objectives of promoting critical thinking, creativity, and comprehensive growth. The effectiveness of machine learning is highly dependent on its careful implementation, similar to any other instrument. The role of technology in education should be seen as an addition to, rather than a replacement for, the intuitive and experiential knowledge possessed by educators.

Numerous scholars have previously underscored the possible drawbacks associated with the convergence of machine learning and the field of education. A variety of challenges can be identified, encompassing ethical considerations such as the possibility of algorithmic prejudice, as well as educational matters that involve an excessive dependence on quantitative measurements to the detriment of qualitative perspectives. This study emphasizes the utilization of data analytics while also advocating for a comprehensive approach that combines algorithmic precision with the profound and multifaceted insights derived from traditional educational methods.

The values of MAE and RMSE are very close in our study, so it might be beneficial to include additional criteria to determine the best model. Some factors to be considered are as follows:

- Computational efficiency: if some models are substantially faster to train or require less memory, this could be a deciding factor.
- Model complexity: simpler models are generally preferable if performance metrics are very close, as they are easier to interpret and less likely to overfit.
- Domain-specific criteria: are there specific requirements in the educational context that might make one type of error (e.g., false positives vs. false negatives) more critical than another?
- Statistical significance: conducting a statistical test to check whether the differences in MAE and RMSE are statistically significant could provide a more definitive answer.

If none of these additional criteria provide a clear winner, it is scientifically acceptable to state that the models perform comparably based on the metrics used.

7. Conclusions

Forecasting the academic achievement of students is a crucial and demanding undertaking in the field of education. The anticipation of academic achievement holds significance in enabling students to take charge of their learning and develop self-regulated learning skills. Additionally, it aids educators in identifying students who may be at risk and mitigating the likelihood of academic failure. In this study, a novel framework is presented to handle the students' learning data. To deal with the unstructured form of data, various feature engineering techniques are utilized to make the data into a meaningful form. After feature engineering, feature selection methods are adopted to obtain the best attributes for both regression and classification tasks. For the regression problem, Lasso and Boruta regressors are used to find the best features; however, for classification problems, RFI and RFI are used to attain the outperforming attributes. Linear regressor, support vector regressor, Random Forest Regressor, and Gradient Boosted Regressors are the ML models

that are employed in the regression part of this study, where Gradient Boost has the least MAE and RMSE of 12.93 and 18.28, respectively, in the case of the Boruta selection method. In the case of classification, the ensemble model, XGBoost, neural network, and Decision Tree models are employed. RFE gives slightly better results. However, the XGBoost model has the most promising results with an accuracy of 78% and precision of 79%.

The impact of this research transcends the academic sphere, providing a foundational blueprint for deploying machine learning techniques in educational settings. The meticulous comparative analysis of diverse models and feature selection techniques, executed within a robust methodological framework, offers actionable insights for educators, policymakers, and edtech developers alike. By identifying optimal strategies for predicting student performance, this study not only advances the scientific understanding of educational analytics but also has immediate, practical applications. Specifically, the insights gleaned can inform targeted interventions, curriculum design, and resource allocation, thereby enhancing the overall educational experience and outcomes. Moreover, this study's rigorous approach to model selection—incorporating statistical significance, relative errors, and domain-specific criteria—sets a precedent for future research, promoting methodological rigor and precision. In a world increasingly driven by data, this research serves as a cornerstone for the responsible and effective utilization of data analytics in education, with the potential to catalyze transformative changes in pedagogical strategies and student engagement.

In future studies, this investigation suggests that there exist multiple potential avenues for future research to be pursued. The scope of this study could be broadened to encompass additional feature selection methodologies and machine learning algorithms. Alternative feature selection methods, including Principal Component Analysis (PCA) and Mutual Information-based Feature Selection (MIFS), as well as diverse machine learning models, such as support vector machine (SVM) and Random Forest, could be employed.

Furthermore, the research could be broadened to encompass additional sets of data. The present investigation solely relied on a singular dataset, thereby warranting further exploration into the efficacy of distinct feature selection methodologies and machine learning models across alternative datasets exhibiting varying attributes. Ultimately, the research could be expanded to encompass a more comprehensive examination of the significance of features and their impact on the efficacy of machine learning algorithms. This may entail exploring the correlations between the chosen predictors and the response variable, as well as scrutinizing the interplay among the selected predictors.

Funding: Research Incentive Fund (RIF) by Zayed University (R22046). APC is funded by Zayed University.

Data Availability Statement: The data is obtained from the public domain. https://analyse.kmi.open.ac.uk/open_dataset#about.

Conflicts of Interest: There is no conflict of interest.

References

1. Siemens, G.J.A. Call for papers of the 1st international conference on learning analytics & knowledge (lak 2011). In Proceedings of the 1st International Conference Learning Analytics & Knowledge, Banff, AL, Canada, 27 February–1 March 2011; Volume 29, p. 2020.
2. Powell, S.; MacNeill, S. *CETIS Analytics Series: Institutional Readiness for Analytics*; CORE: Milton Keynes, UK, 2012.
3. Natek, S.; Zwilling, M. Student data mining solution–knowledge management system related to higher education institutions. *Expert Syst. Appl.* **2014**, *41*, 6400–6407. [CrossRef]
4. Kumar, A.D.; Selvam, R.P.; Kumar, K.S. Review on prediction algorithms in educational data mining. *Int. J. Pure Appl. Math.* **2018**, *118*, 531–537.
5. Liu, Q.; Wu, R.; Chen, E.; Xu, G.; Su, Y.; Chen, Z.; Hu, G. Fuzzy cognitive diagnosis for modelling examinee performance. *ACM Trans. Intell. Syst. Technol.* **2018**, *9*, 1–26. [CrossRef]
6. Fausett, L.; Elwasif, W. Predicting performance from test scores using backpropagation and counterpropagation. In Proceedings of the 1994 IEEE International Conference on Neural Networks (ICNN'94), Orlando, FL, USA, 28 June–2 July 1994; pp. 3398–3402.

7. Livieris, I.E.; Drakopoulou, K.; Mikropoulos, T.A.; Tampakas, V.; Pintelas, P. An ensemble-based semi-supervised approach for predicting students' performance. In *Research on e-Learning and ICT in Education*; Technological, P., Perspectives, I., Eds.; Springer: Cham, Switzerland, 2018; pp. 25–42.
8. Loh, C.S.; Sheng, Y.J.E. Measuring the (dis-) similarity between expert and novice behaviors as serious games analytics. *Educ. Inf. Technol.* **2015**, *20*, 5–19. [CrossRef]
9. Wook, M.; Yusof, Z.M.; Nazri, M.Z.A. Educational data mining acceptance among undergraduate students. *Educ. Inf. Technol.* **2017**, *22*, 1195–1216. [CrossRef]
10. Picciano, A.G. The evolution of big data and learning analytics in American higher education. *J. Asynchronous Learn. Netw.* **2012**, *16*, 9–20. [CrossRef]
11. Viberg, O.; Hatakka, M.; Bälter, O.; Mavroudi, A. The current landscape of learning analytics in higher education. *Comput. Hum. Behav.* **2018**, *89*, 98–110. [CrossRef]
12. Kotsiantis, S.B.; Pierrakeas, C.; Pintelas, P.E. Preventing student dropout in distance learning using machine learning techniques. In Proceedings of the Knowledge-Based Intelligent Information and Engineering Systems, 7th International Conference, Oxford, UK, 3–5 September 2003; pp. 267–274.
13. Romero, C.; Ventura, S. Educational data mining: A survey from 1995 to 2005. *Expert Syst. Appl.* **2007**, *33*, 135–146. [CrossRef]
14. Romero, C.; Ventura, S. Educational data mining: A review of the state of the art. *IEEE Trans. Syst. Man Cybern. Part C* **2010**, *40*, 601–618. [CrossRef]
15. Minaei-Bidgoli, B.; Kashy, D.A.; Kortemeyer, G.; Punch, W.F. Predicting student performance: An application of data mining methods with an educational web-based system. In Proceedings of the 33rd Annual Frontiers in Education, 2003 (FIE 2003), Westminster, CO, USA, 5–8 November 2003; p. T2A-13.
16. Peña-Ayala, A. Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Syst. Appl.* **2014**, *41*, 1432–1462. [CrossRef]
17. Shih, B.-Y.; Lee, W.-I. The application of nearest neighbor algorithm on creating an adaptive on-line learning system. In Proceedings of the 31st Annual Frontiers in Education Conference—Impact on Engineering and Science Education—Conference Proceedings (Cat. No. 01CH37193), Reno, NV, USA, 10–13 October 2001; p. T3F-10.
18. Younas, J.; Lukowicz, P. Cognitive Ability Classification using On-body Sensors. In Proceedings of the Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers, Cambridge, UK, 11–15 September 2022; pp. 317–320.
19. Kuzilek, J.; Hlosta, M.; Herrmannova, D.; Zdrahal, Z.; Vaclavek, J.; Wolff, A. OU Analyse: Analysing at-risk students at The Open University. *Learn. Anal. Rev.* **2015**, *LAK15-1*, 1–16.
20. He, J.; Bailey, J.; Rubinstein, B.; Zhang, R. Identifying at-risk students in massive open online courses. In Proceedings of the AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015.
21. Kovacic, Z. Early prediction of student success: Mining students' enrolment data. In Proceedings of the InSITE 2010: Informing Science + IT Education Conference, Cassino, Italy, 19–24 June 2010.
22. Kotsiantis, S.; Patriarcheas, K.; Xenos, M. A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education. *Knowl. Based Syst.* **2010**, *23*, 529–535. [CrossRef]
23. Osmanbegovic, E.; Suljic, M. Data mining approach for predicting student performance. *Econ. Rev.* **2012**, *10*, 3–12.
24. Watson, C.; Li, F.W.; Godwin, J.L. Predicting performance in an introductory programming course by logging and analyzing student programming behavior. In Proceedings of the 2013 IEEE 13th international conference on advanced learning technologies, Beijing, China, 15–18 July 2013; pp. 319–323.
25. Hu, Y.-H.; Lo, C.-L.; Shih, S.-P. Developing early warning systems to predict students' online learning performance. *Comput. Hum. Behav.* **2014**, *36*, 469–478. [CrossRef]
26. Lakkaraju, H.; Aguiar, E.; Shan, C.; Miller, D.; Bhanpuri, N.; Ghani, R.; Addison, K.L. A machine learning framework to identify students at risk of adverse academic outcomes. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, 10–13 August 2015; pp. 1909–1918.
27. Ahmed, A.; Elaraby, I.S. Data mining: A prediction for student's performance using classification method. *Int. J. Comput. Sci. Eng.* **2014**, *2*, 43–47. [CrossRef]
28. Marbouti, F.; Diefes-Dux, H.A.; Madhavan, K. Models for early prediction of at-risk students in a course using standards-based grading. *Comput. Educ.* **2016**, *103*, 1–15. [CrossRef]
29. Iqbal, Z.; Qadir, J.; Mian, A.N.; Kamiran, F. Machine learning based student grade prediction: A case study. *arXiv* **2017**, arXiv:1708.08744.
30. Almarabeh, H. Analysis of students' performance by using different data mining classifiers. *Int. J. Mod. Educ. Comput. Sci.* **2017**, *9*, 9–15. [CrossRef]
31. Xu, J.; Moon, K.H.; Van Der Schaar, M. A machine learning approach for tracking and predicting student performance in degree programs. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 742–753. [CrossRef]
32. Al-Shehri, H.; Al-Qarni, A.; Al-Saati, L.; Batoaq, A.; Badukhen, H.; Alrashed, S.; Alhiyafi, J.; Olatunji, S.O. Student performance prediction using support vector machine and k-nearest neighbor. In Proceedings of the 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), Windsor, ON, Canada, 30 April–3 May 2017; pp. 1–4.

33. Daud, A.; Aljohani, N.R.; Abbasi, R.A.; Lytras, M.D.; Abbas, F.; Alowibdi, J.S. Predicting student performance using advanced learning analytics. In Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, 3–7 April 2017; pp. 415–421.
34. Masci, C.; Johnes, G.; Agasisti, T. Student and school performance across countries: A machine learning approach. *Eur. J. Oper. Res.* **2018**, *269*, 1072–1085. [CrossRef]
35. Aggarwal, D.; Mittal, S.; Bali, V. Significance of non-academic parameters for predicting student performance using ensemble learning techniques. *Int. J. Syst. Dyn. Appl.* **2021**, *10*, 38–49. [CrossRef]
36. Zeineddine, H.; Braendle, U.; Farah, A.J.C.; Engineering, E. Enhancing prediction of student success: Automated machine learning approach. *Comput. Electr. Eng.* **2021**, *89*, 106903. [CrossRef]
37. Buenaño-Fernández, D.; Gil, D.; Luján-Mora, S.J.S. Application of machine learning in predicting performance for computer engineering students: A case study. *Sustainability* **2019**, *11*, 2833. [CrossRef]
38. Hussain, M.; Zhu, W.; Zhang, W.; Abidi, S.M.R.; Ali, S. Using machine learning to predict student difficulties from learning session data. *Artif. Intell. Rev.* **2019**, *52*, 381–407. [CrossRef]
39. Alhusban, S.; Shatnawi, M.; Yasin, M.B.; Hmeidi, I. Measuring and enhancing the performance of undergraduate student using machine learning tools. In Proceedings of the 2020 11th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 7–9 April 2020; pp. 261–265.
40. Yukselturk, E.; Ozekes, S.; Turel, Y.K. Predicting dropout student: An application of data mining methods in an online education program. *Eur. J. Open Distance E-Learn.* **2014**, *17*, 118–133. [CrossRef]
41. Wang, W.; Yu, H.; Miao, C. Deep model for dropout prediction in MOOCs. In Proceedings of the 2nd International Conference on Crowd Science and Engineering, Beijing, China, 6–9 July 2017; pp. 26–32.
42. Aulck, L.; Velagapudi, N.; Blumenstock, J.; West, J. Predicting student dropout in higher education. *arXiv* **2016**, arXiv:1606.06364.
43. Thaker, K.; Huang, Y.; Brusilovsky, P.; Daqing, H. Dynamic knowledge modeling with heterogeneous activities for adaptive textbooks. In Proceedings of the 11th International Conference on Educational Data Mining, Buffalo, NY, USA, 15–18 July 2018; pp. 592–595.
44. Ahadi, A.; Lister, R.; Haapala, H.; Vihavainen, A. Exploring machine learning methods to automatically identify students in need of assistance. In Proceedings of the Eleventh Annual International Conference on International Computing Education Research, Omaha, NE, USA, 9–13 July 2015; pp. 121–130.
45. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* **1996**, *58*, 267–288. [CrossRef]
46. Kursa, M.B.; Rudnicki, W.R. Feature selection with the Boruta package. *J. Stat. Softw.* **2010**, *36*, 1–13. [CrossRef]
47. Genuer, R.; Poggi, J.-M.; Tuleau-Malot, C. Variable selection using random forests. *Pattern Recognit. Lett.* **2010**, *31*, 2225–2236. [CrossRef]
48. Granitto, P.M.; Furlanello, C.; Biasioli, F.; Gasperi, F. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemom. Intell. Lab. Syst.* **2006**, *83*, 83–90. [CrossRef]