sustainability

MDPI

Article

An Interpretable Framework for an Efficient Analysis of Students' Academic Performance

Ilie Gligorea [1,2,*], Muhammad Usman Yaseen [3,*], Marius Cioca [4], Hortensia Gorski [1] and Romana Oancea [1]

1 Department of Technical Science, Faculty of Military Management, "Nicolae Bălcescu" Land Forces Academy, 550170 Sibiu, Romania; hortensia.gorski@gmail.com (H.G.); oancea.romana@gmail.com (R.O.)
2 Doctoral School, University of Petroșani, 332006 Petroșani, Romania
3 Computer Science Department, Comsats University, Islamabad 45550, Pakistan
4 Department of Industrial Engineering and Management, Faculty of Engineering, "Lucian Blaga" University of Sibiu, 550025 Sibiu, Romania; marius.cioca@ulbsibiu.ro
* Corespondence: iliegligorea@gmail.com (I.G.); muhammadusmanyaseen@gmail.com (M.U.Y.)

Abstract: Recent technological advancements in e-learning platforms have made it easy to store and manage students' related data, such as personal details, initial grade, intermediate grades, final grades, and many other parameters. These data can be efficiently processed and analyzed by intelligent techniques and algorithms to generate useful insights into the students' performance, such as to identify the factors impacting the progress of successful students or the performance of the students who are struggling in their courses and are at risk of failing. Such a framework is scarce in the current literature. This study proposes an interpretable framework to generate useful insights from the data produced by e-learning platforms using machine learning algorithms. The proposed framework incorporates predictive models, as well as regression and classification models to analyze multiple factors of student performance. Classification models are used to systematize normal and at-risk students based on their academic performance, with high precision and accuracy. Regression analysis is performed to determine the inherent linear and nonlinear relationships between the academic outcomes of the students acting as the target or independent variables and the performance indicative features acting as dependent variables. For further analysis, a predictive modeling problem is considered, where the performance of the students is anticipated based on their commitment to a specific course, their performance for the whole course, and their final grades. The efficiency of the proposed framework is also optimized by reliably tuning the algorithmic parameters. Furthermore, the performance is accelerated by empowering the system with a GPU-based infrastructure. Results reveal that the proposed interpretable framework is highly accurate and precise and can identify factors that play a vital role in the students' success or failure.

Keywords: training system; e-learning platform; machine learning; regression

Citation: Gligorea, I.; Yaseen, M.U.; Cioca, M.; Gorski, H.; Oancea, R. An Interpretable Framework for an Efficient Analysis of Students' Academic Performance. *Sustainability* **2022**, *14*, 8885. https://doi.org/10.3390/su14148885

Academic Editors: Dimitrios Stamovlasis, Michail Kalogiannakis and Theodosios Sapounidis

Received: 16 May 2022
Accepted: 18 July 2022
Published: 20 July 2022

1. Introduction

Advancements in the computing domain, including cloud computing, fog computing, and edge computing, and software-related domains, such as big data technologies, machine learning, and algorithms of artificial intelligence, have revolutionized the world. Most of the tasks that were difficult or impossible to achieve a few years ago can now be performed in no time. Technological advancements have also transformed not only the business sector, but also the educational and academic sectors [1,2]. In recent years, higher education institutions (HEIs) have become rapidly developing sectors, due to the utilization of these advanced technologies. Furthermore, since HEIs are committed to achieving, creating, disseminating, and publishing knowledge, adapting to the ever-growing new techniques and technologies has become a vital aspect [3]. HEIs can adapt and harness the power of these technologies to improve the quality of research and teaching practices [4]. It has become obvious that the number of students enrolled in HEIs is constantly increasing. On

the one hand, this fact is encouraging because it contributes to the growth of the literacy rate of the society; on the other hand, it has limitations that need to be addressed. One major reason for the increase in the number of enrollments per year is the financial needs of the institutions. Meeting the requirements of the students and providing all the students with an active interaction within the faculty requires labor and human resources. Institutions find it difficult to cope with these escalating requirements for resources and are forced to increase their students' possibility to generate more revenue. Another major reason for the increased student strength in the education sector is the growing emphasis on offering students opportunities for higher education, for better careers and professional development. The dramatic increase in the number of students enrolled per year generates numerous challenges, such as monitoring all students, class sizes, faculty–student interaction, research support for students, collaboration between students and faculty on research projects, skill enhancement, and grading [5].

The traditional and classical approach of manually handling and managing these challenges has now become obsolete. This approach fails to provide active feedback to students and to pressure students for timely completion of tasks [6]. In order to respond to the above-mentioned challenges, HEIs are now trying to adapt to the recent technological advancements by implementing better evaluation, assessment, and managerial solutions to show their performance, efficiency, and productivity.

Another major problem posed by increased student strength in HEIs is continuous monitoring of student learning and progress. Efficiently tracking the progress and learning outcomes of a student is a hectic and time-consuming task. E-learning platforms and learning management systems (LMSs) can help cope with most of these challenges and are becoming a necessity in order to meet the demands of the modern education system [7,8]. These platforms not only help to manage the activities of a large number of students, but can also store and retain data that can also be used to perform learning analytics, such as student performance in multiple subjects, students at risk, and the factors involved in student success [9]. Even a simple variable such as the log time of students in the LMSs can help better predict the overall performance of a student. Studies have also shown that there is a direct connection between the success of a student gaining good marks and the number of times a student has accessed the management system or platform. Recent studies have identified a direct connection between these two variables [10,11].

E-learning and management platforms assist teachers and faculty members and can also help students to predict early signs of their performance at the end of the term period. Consequently, they can act upon obtaining the necessary measures to improve their performance and ultimately have a better outcome. With a fusion of multiple domains including cloud computing, fog and edge computing, artificial intelligence, and machine learning, it is possible to build powerful LMSs and portals that can help perform all the above-mentioned tasks. The most recent learning management systems are web- and cloud-based platforms that offer multiple services, such as student attendance mechanisms, online submission of answer sheets and their grading, online availability of lecture slides, notes, and other related materials, online submission and assessment tools for quizzes, assignments, and final exams, and communication and collaboration tools for interaction with faculty and among students [12].

Academic analytics is the process of analyzing the data generated by these platforms to produce useful insights for faculty, as well as for students [13]. The term academic analytics is originated in the term 'business analytics' and has gained worldwide popularity in recent years. However, in the education sector, a similar approach is taken in which academic-related data generated by LMSs and online portals are processed by intelligent and statistical algorithms backed by powerful infrastructures, such as cloud and graphical processing units (GPUs) [14]. So far, these approaches are mostly used to process data collected at the institution level to observe student demographics and to monitor their overall performance in terms of the institution's ranking and its score among other institutions.

However, there are very limited studies using these technological advancements to address problems such as student retention, student at-risk cases, and enhanced learning practices.

In this study, we propose a machine-learning-based interpretable framework for e-learning platforms with the basic aim of helping both students and teachers improve their performance. The proposed framework incorporates predictive models, as well as regression and classification models to analyze multiple activities of students. The research question this study is trying to answer is whether it is possible to incorporate predictive models, as well as regression and classification models in an interpretable framework to generate useful insights by using the students' data produced by e-learning platforms. Classification models are used to rank students based on their performance. For instance, a classification model is designed and developed to sort normal and at-risk students based on their academic performance with high precision and accuracy. The regression analysis is performed to determine the inherent linear and nonlinear relationships between the academic outcomes of the students acting as the target or independent variable and the performance indicative features acting as dependent variables.

A predictive modeling problem is also considered, in which students' performance is anticipated based on their commitment to the courses. Predictive models also help speculate about a possible problematic situation that any student could raise during their learning processes. These models also attempt to explain what has happened specifically in a particular situation without knowing the intention of the target in advance. These algorithms are used in our framework to perform an analysis of the data produced by e-learning platforms to help improve student and teacher performance. The efficiency of the proposed framework is also optimized by successfully tuning the algorithmic parameters, and the performance is accelerated by empowering the system with a GPU-based infrastructure.

The sections of the paper are organized as follows. Section 2 describes the related work; Section 3 highlights the role of machine learning in academic analytics and how it is being used to generate analytics for e-learning platforms; Section 4 details the proposed machine-learning-based framework for academic analytics and explains the implementation details; Section 5 presents the experimental setup, while the generated results are detailed in Section 6; Section 7 concludes the paper with future avenues for research.

2. Related Work

Academic analytics has been an active area of research in the past decade. This area has gained even more attention in the last couple of years due to the COVID-19 pandemic. Most academic activities are now being held online, which generates a large amount of data. Learning management systems and e-portals are playing an essential role in managing and storing these data. Cloud computing technologies and hardware accelerators such as GPUs also help LMSs and e-portals to efficiently perform the storage and analysis tasks. The recent studies surveyed in this section have made an essential contribution to the process of analyzing these data with the help of statistical techniques and machine learning algorithms.

In the past, educational data mining (EDM) research was a focus of many researchers [15–17] who analyzed LMS-generated data containing information such as log information, activity records, and progress details. The basic objective of EDM is to generate useful insights from the data, such as identification of at-risk students [18], estimating the efficiency of a learning system [19], understanding the behavior of successful and struggling students [20], etc.

Studies have also been conducted to help not only students, but also faculty, teachers, and non-teaching staff by developing warning systems that could predict the possible outcomes of struggling students [21]. These studies also suggest possible measures and actions that could be taken to move these students from the failure stage to the success track. Decision support systems [22] have also been cultivated to provide instructors with mechanisms that help guide struggling students.

According to a recent survey [23], EDM studies have mainly incorporated issues based on classification [24] or clustering [25] in the setting of online e-portals and platforms. Predictive analytics [26] has also been performed, and the main focus was not on the hypothesis, but rather on the data characteristics, such as predicting the outcome of an assessment performed in an online setting [27], proposing and developing a platform capable of predicting the outcome of a test in a timely and efficient manner. Similarly, in another study [28], an online system was developed for distance-based learning, and the objective was to predict the performance for a particular test or exam.

Classification techniques of machine learning have been employed to rank the factors involved in the performance of both successful students and unsuccessful students [29]. An effort has also been made to classify the common student characteristics and behaviors that drive them to perform better on a test or exam [30]. These identified characteristics can provide an association with the future success of the students.

Another category is the use of classical machine learning algorithms, such as naive Bayes and decision trees, to perform prediction or classification. These predictions or classifications are made based on a feature set that is mainly hand-crafted. These hand-crafted feature sets were provided to the classifier to make predictions or classifications. One such study [31] used decision trees (DT) to help students make learning better in an e-learning setting. Similarly, naive Bayes, fuzzy logic [32], and fuzzy inductive systems and reasoning [33] were also used to construct classification and prediction models.

Artificial neural networks [34], which are very famous classification algorithms and constitute the basis of modern deep learning algorithms, have also played a role in academic analytics. They consist of multiple layers, and each layer is made up of several neurons. Usually, there are three layers present in a traditional neural network. One is the input layer, which receives the input; the second is the hidden layer; the third is the output layer, which generates the output. These were mainly used to establish classifiers that can predict final grades, organize behavioral patterns, and identify personality traits and other numerous characteristics.

It is particularly important to highlight that all the above-mentioned studies have developed classifiers with variables that were not directly time-bound and had no influence on time-series data. The main reason behind this was the use of a traditional classroom-based approach to conducting teaching and other related activities. Learning management systems just played an assisting role in all these studies [35]. However, some of the above-cited studies have been very successful in shedding light on at-risk students or in generating early warnings under constrained conditions.

Most of the studies cited above have performed statistical analyses and some of them have used machine learning approaches. This revealed the fact that machine-learning-enabled learning management systems and e-portals are not very mature yet and are still in their infancy period. Moreover, these approaches have targeted very specialized areas, such as the success or failure rate of students, but these approaches still fail to identify more generic aspects of the underlying problems, such as the classification of factors involved in students' failure or success or understanding the personality traits of a student, which represent an obstacle in student success, and the overall performance characterization of students.

Most of the studies on machine-learning-enabled learning management systems also remain unable to fully exploit the potential of machine learning techniques to improve the quality and performance of existing LMSs and e-portals [36]. These studies are still far from adding intelligence to existing systems and cannot be directly used with recent technologies such as virtual reality (VR) and augmented reality (AR). Hence, these systems still need much work and effort to meet all the challenges of traditional learning management systems.

Although state-of-the-art approaches that utilize data generated by e-portals or LMSs do very well in performing academic analytics, there are a few significant limitations. Most of the approaches are specifically speculated to generate analytics in a single context and

cannot be generalized. They have models trained on a limited dataset and are one-offs. They seriously lack generalization capacity and, thus, cannot be extended in functionality for multiple uses. Moreover, in most cases, a model designed and trained for a particular subject, course, or institution is not feasible and cannot be expected to be applied to other domains or courses. There is a pertinent need to develop a model that could adopt a considerable amount of variability in different contexts and could be practiced across courses and multiple institutions.

Another major limitation is that current approaches act as a black box and it is very difficult, sometimes even impossible to interpret the results. This is independent of the fact that the model performs prediction or classification or is a descriptive model. This hinders the process of final decision-making in relation to the progress of students or institutions. The alert or warning signals generated by the system need to be interpreted in a meaningful way so that necessary measures can be taken at a specific moment, which will capitalize on the true benefit of academic analytics.

3. Machine Learning in Academic Analytics

Machine learning has transformed the world in almost all domains and is also playing a vital role in academic analytics. Machine learning has its applications ranging from medical engineering to aerospace and is contributing to enhancing and improving the functionality, performance, and accuracy of the systems involved in all disciplines. In academic analytics, researchers are using machine learning to improve the performance and accuracy of their learning management systems and e-learning platforms [37]. Algorithms have been developed and new approaches have been proposed to perform regression, classification, and prediction in LMSs and e-learning portals. These algorithms not only enhance the functionality of the learning systems, but also make them adaptive with more generalization capability.

Machine learning algorithms are hungry for data, and the learning management system is full of data that can be retained for a very long period. More specifically, if the LMS is backed by a cloud storage platform, then retaining and processing data becomes quite easy. This leads to the usage of machine learning algorithms for the processing of these data to generate academic analytics [38]. The data stored in the cloud platform by these LMSs can be in the form of log files, course materials, examination materials and their results, demographic information related to students, and other specific details. These data can be processed by machine learning algorithms to produce useful insights and analytics.

It is pertinent to mention here that the existing machine-learning-based approaches have developed classifiers with variables that are not directly time-bound and have no influence on time-series data. Furthermore, the scope of these approaches was very limited, and they were designed to target very specialized areas, such as the success or failure rate of students, revealing the fact that machine-learning-enabled learning management systems and e-portals are not very mature yet and are still in their infancy period. Moreover, existing studies on machine-learning-enabled learning management systems also remained unable to fully exploit the potential of machine learning techniques to improve the quality and performance of existing LMSs and e-portals. Most of the analysis performed in these studies is unidirectional, which means that they are either working on the regression analysis by consuming some dependent or independent variables or perform classification based on the grades of the students. These approaches fail to identify more generic aspects of the underlying problems, such as the classification of factors involved in students' failure or success or understanding the personality traits of a student, which represent an obstacle to student success, and the overall performance characterization of students.

Another major limitation is that current approaches mainly use neural networks trained on CPU-based infrastructure without any optimization strategy, which acts as a black box and, accordingly, it is very difficult, sometimes even impossible to interpret the results. This is independent of the fact that the model performs prediction or classification or is a descriptive model. This hinders the process of final decision-making in terms of the

progress of students or institutions. The alert or warning signals generated by the system need to be interpreted in a meaningful way so that necessary measures can be taken at a specific moment, which will capitalize on the true benefit of academic analytics. Hence, these systems still need much work and effort to meet all the challenges of traditional learning management systems.

3.1. Academic Analytics and Classification

Classification algorithms have been used recently to perform different types of ranking tasks in academic analytics, such as to organize normal and at-risk students or to label the behaviors and patterns common to successful or unsuccessful students. Some researchers have also used classification algorithms to arrange the rates of subject completion by different students and also to predict learning styles in e-learning systems [39]. This was performed using the preferences provided by each student along with other details.

The most commonly used classification algorithms in academic analytics include Bayes networks, decision trees, random forests, support vector machines, and neural networks. These approaches have also been used in abductive network modeling, where the main objective was to classify examinees with high accuracy rates [40]. These approaches have also been used to rank the factors involved in dropout cases. The main focus was to highlight the main causes and classify the issues involved in unsuccessful performance that could cause dropout. These factors are also used in predictive algorithms to anticipate which students may potentially experience dropout [41].

Decision trees and support-vector-machine-based approaches [42,43] have been employed in e-learning courses to develop classification models for optimizing learning sequences. Student profile data and their progress details over a limited period are used in decision trees to identify successful student trait sequences for a particular subject [44]. To optimize the process of adaptive learning sequences, decision trees have been adapted to improve accuracy and reduce the classification time. Similarly, support vector machines tend to classify patterns and behaviors of both successful and struggling students based on their support vectors. A decision boundary is drawn with the help of selected support vectors to identify patterns and pave a path for student success based on these patterns.

Neural networks have their applications mainly in web-based learning management systems and still need to be utilized to their full extent and power. These are powerful algorithms, but are still under-utilized in this domain. Neural networks are mostly used to identify and extract learning patterns that play a major role in student success [45]. These extracted learning patterns can be used by other struggling students to improve learning and maximize training. Some studies have also used neural networks to predict and identify a subject for a particular student based on his/her history and performance [39].

3.2. Academic Analytics and Regression

In addition to classification algorithms, regression analysis has also been used in academic analytics to improve the performance of e-learning and learning management systems [46]. There are two types of regression algorithms, both of which have been utilized to produce analytics in the education sector. The former is a linear regression, which is mostly used when there is a binary case of two possible variables and the value of one variable depends on the value of the second variable, i.e., one is dependent, and the other is an independent variable. This much simpler case is used when the number of factors involved in the learning patterns or observations is small [47]. The latter is a logistic regression which is exploitable when there are many available factors that contribute to the failure or success of a student.

To perform a regression analysis, it is relevant to have data that are time-series in nature. For this reason, the most suitable data to perform academic analysis in LMSs are the online activity data of the students. An important variable in the online activity data is the student login information on the e-learning portal or platform [48]. It has been observed

that the login information, along with other parameters, such as test scores and attendance, provides very good results for the prediction of future student success.

Regression analysis has also been performed on the LMS tracking data to identify the variables that have a direct correlation with the outcome of the student. A single course-based regression analysis has been performed in most studies to produce a model that could provide best-fit predictions [49]. These best-fit predictive models have identified the major factors that play an important role in higher accuracy and performance, such as login information, number of posts, email communication, total assessments, etc.

3.3. Association Rules and Clustering in Academic Analytics

Association rules refer to a classical machine learning approach that is mostly used to find similar patterns and behaviors and discover commonalities between the underlying datasets. In the context of learning management systems and e-portals, association rules have been used to identify student behavior and misconceptions, which are common to struggling students [50]. One of the most-used association rule algorithms, known as the a priori algorithm, has been used for this task, and it can help to dig into the profile of students who have been unsuccessful in a particular course [51]. The information generated by these algorithms can help the teachers and faculty to take the necessary actions and measures for the improvement of the students performance.

Clustering is an unsupervised machine learning approach that aims to create clusters of correlated samples present in the datasets. It is only utilized when the labeled training and testing data are not available and, thus, a grouping is performed based on the similar characteristics of the dataset samples. A distance measure is employed to determine the distance between each sample of the dataset. Samples are assigned to groups or clusters based on the distance value.

Clustering in academic data analytics is used to create groups or clusters of students that have similar learning patterns and some common traits. It is also exploited to group common learning profiles of different students so that a learning policy could be established and applied to particular groups for further improvement [52]. Clustering in learning management systems and e-portals can also help form groups of successful or struggling students when there is no labeling provided by the portals or management systems [53].

Deep learning, which refers to advancement of classical machine learning algorithms, especially neural networks, has proven to be very successful in almost all domains. It is an advanced form of the conventional neural network in which the depth of the network is increased by adding multiple operations such as convolution, pooling, drop out, etc., according to the requirements of the application. These are also useful in generating analytics from different types of datasets such as images, videos, text, or numerical data. Some recent studies have also applied deep learning algorithms in the field of academic analytics [54,55], but there are still gaps that need to be filled in this area of research.

4. Proposed Framework and Implementation Details

This section describes the proposed machine-learning-based framework and explains each of its components in detail. The flow of the framework is as follows: the machine learning models were trained under the supervised learning domain to perform classification and regression analysis. However, before using the machine learning network, the underlying data were preprocessed, cleaned, and then, separated into training and testing phases. The training phase extracts useful features from the data to generate analytics. Both shallow and deep learning networks were experimented with, in order to see which of them provides the best analytical results.

The proposed machine learning pipeline for academic analytics is illustrated in Figure 1. As it can be seen from the figure, the dataset is first loaded into the system memory using the Pandas library to perform data preprocessing and cleaning. After performing all preprocessing steps, the dataset was split into a training and test set. For this purpose, the Sklearn library was used. After that, the system training was performed with

highly tuned hyperparameters, including regularization, learning rate, and number of branches to generate analytics.

The proposed machine-learning-based interpretable framework consists of several vital components working in a cascaded fashion from initial preprocessing of the raw data to final prediction and classification. The output of each preceding component becomes the input of the following component to generate useful insights from the data produced by e-learning platforms. The first component as shown in Figure 1 prepares the data and applies multiple preprocessing steps to them to make feasible model training. The preprocessing phase ensures that there are no missing values and resolves any class imbalance issues.

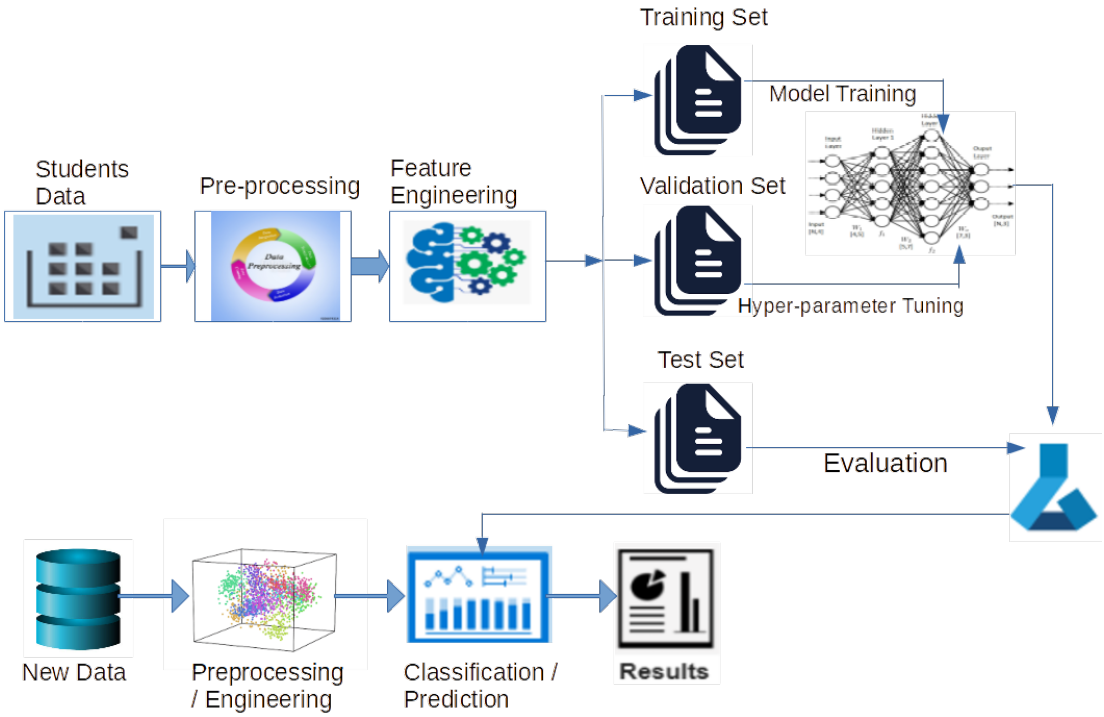

Figure 1. Machine-learning-based interpretable framework for e-learning platforms.

Another major component of the proposed framework is feature engineering in which the most optimal features are engineered based on the attributes required for the specified task. For this, the data are encoded to convert the categorical values to numerical values and then scaled to standardize them in a fixed range. The features are then engineered on the basis of a specific task. For instance, the assessment features describe the knowledge of the students for each course, which can be engineered based on the grades of the students. This can be helpful for a prediction case. Similarly, the status and participation of the student are also important factors and were engineered together for an accurate prediction.

The proposed framework was further empowered by incorporating predictive models, as well as regression and classification models to analyze multiple factors of students' performance. Classification models systematize normal and at-risk students based on their academic performance, while regression analysis was performed to determine the inherent linear and nonlinear relationships between the students' academic outcomes acting as independent variables and the performance indicative features acting as dependent variables. The framework also incorporates predictive modeling in which the performance of the students is anticipated based on their commitment to a specific course, their performance for the whole course, and their final grades. Each component of the proposed machine-learning-based interpretable framework is detailed in the coming subsections.

4.1. Data Preparation and Pre-Processing

A publicly available or self-generated dataset often contains a number of challenges. These challenges are usually handled in the preprocessing stage and they mainly include missing values, class imbalance, merging and grouping different tables of the dataset if necessary, and fixing data types and inconsistent weights.

The open university dataset used in this study mainly consists of CSV files. The data from these CSV files were loaded into the data frames and, then, pre-processing was applied. For certain cases, different data tables were merged, and the columns were reordered. Grouping was also performed so that a prediction can be made for each user. Moreover the mean was employed to impute the missing values in most of the cases.

4.2. Data Encoding

Most machine learning algorithms cannot handle categorical variables unless they are converted to numerical values. Many of the algorithm performances vary based on how categorical variables are encoded. For this reason, different methods for encoding categorical variables were employed, including:

- OrdinalEncoder to encode binary categorical variables;
- OneHotEncoder to encode nominal categorical variables;
- StandardScaler to standardize numerical variables.

4.3. Data Scaling

Data scaling is an important part of machine learning algorithms. The features to be found in any dataset can be independent and, therefore, it is necessary to standardize them so that they can be in a fixed range. If scaling is not performed before training the machine learning model, then the model will weigh larger values higher and smaller values lower, which could affect the overall performance of the model. MinMaxScaler and StandardScaler were used for this purpose. The features were also standardized by removing the mean and dividing by the variance.

4.4. Feature Engineering

To tackle the problem as a performance prediction problem, optimal feature engineering was performed on the available dataset based on the attributes required for this task. The assessment of the students mentioned in the assessment table describes the knowledge of the students for each course. It also contains the grades of the students, which can be very helpful for a prediction case.

However, the structure of the dataset is distinct for different courses, which needs to be addressed efficiently. Therefore, the final grade of the student was engineered with the weight of his/her assessment, and the pass rate based on the assessment of the student was also used. The status and participation of the student are also important factors and were engineered together for an accurate prediction.

Another essential element is the data inferred from the virtual learning environment, which explain the interaction of the students with the course materials available to them during the whole course of study. This information was an indicator used to assess student contact hours with the subjects and how well they studied them.

4.5. Exploratory Data Analysis

Different visualization tools and techniques have been utilized to perform exploratory data analysis to gain a better understanding of the data. This is done with various visualization and plotting functions such as heat maps, histograms, count plots, and box plots. These visualizations and plotting functions help to better understand the relationships between different features present in the dataset. The categorical feature proportion can be easily visualized through the count plot and each bar percentage can be annotated over it. The resulting numerical data can be considered categorical data or ordinal categorical data and can be represented by numbers.

4.6. Stratified Sampling

Stratified sampling was also employed in this work. Since the size of the dataset appears to be small, random sampling is not a good approach. The better approach to be followed here is stratified sampling. Therefore, stratified sampling was implemented by selecting a main characteristic and separating the dataset into multiple strata. For cross-validation, K-fold cross-validation and stratified K-fold were used. Training and testing data were separated as a percentage of 80 and 20, respectively [56].

4.7. Feature Selection and Reduction

A feature reduction technique was used to select important features and reduce the dimensions of the features. Principal component analysis was applied to select the most important and distinguishing features of the input dataset. The selection and reduction of features are important processes to optimize the working of a system. A system without a feature selection and reduction process will attempt to employ all features present in the dataset.

Most of these features will not be useful and will not play any role in the performance of the system. Therefore, this will result in an increase in the computation cost, as well as in memory space. A system employing a feature selection or reduction mechanism will not only reduce the computation cost, but will also save training time and memory.

4.8. Regression Analysis

Different machine learning algorithms were experimented with in order to generate analytics from the data. Both regression and classification analyses were used. The following machine learning models were used mainly to perform regression, which helped generate insights from the data. For regression analysis, linear regression, logistic regression, lasso regression, and support vector regression were applied.

Before performing the regression analysis, some tables presented in the dataset were merged, as there was some missing information. Moreover, there was information required from multiple tables. For example, to pass a course, the student needs to be successful in both assignments and the exam. Therefore, for this purpose, some tables needed to be merged as per the requirements. For regression analysis, all tables were merged to create a single dataset. However, some unnecessary columns were dropped, and the dataset was cleaned.

4.9. Classification

As mentioned above, both regression analysis and classification were used to generate the required insights from the data. The following machine learning models were mainly used to perform classification: decision tree, random forest, and support vector machines. All preprocessing steps that were carried out for regression analysis were also performed for classification. This included merging the required tables and dealing with the missing information. The categorical values of the dataset were encoded as previously explained, and then, the classification models were trained and evaluated. The unnecessary columns were dropped, and the datasets were cleaned.

4.10. System Optimization

We optimized the classifiers used in this work with the stochastic gradient descent classifiers. Stochastic gradient descent (SGD) is an optimization model that helps to minimize or maximize a loss function depending on the application. However, most of the time, it is used to minimize the cost function. In our scenario, SGD helped to perform mini-batch learning. The use of SGD was of utmost importance, especially for support vector machines, because it was difficult to calculate the cost function directly.

5. Experimental Setup

This section describes the experimental setup of our proposed system. First, the dataset in this study is outlined with proper visualization to understand the dataset and to determine the relationships between different features and attributes. Subsequently, we detail the tools and technologies used to implement the proposed system and then explain the performance metrics used to measure the performance of the system.

5.1. Dataset

The dataset used in this work is an open-source dataset that depicts student behavior in an open university [57]. The dataset provides critical and very useful information about the behavior and interactions of the students. It provides all sorts of student information and other data related to student relations, which were collected from a learning management system in a virtual learning environment. The data attributes include student grades, demographic logs, and assessment-related features.

The uniqueness of the dataset lies in the fact that it consists of demographic data and clickstream information from students. The clickstream information depicts the interaction of students with the learning environment. This information proved to be very helpful in analyzing the behavior of the students based on the actions they performed. There were around 10,655,280 clicks from 32,593 students enrolled in 22 courses along with their assessment results. Another advantage of exploiting this dataset is that the dataset is in tabular format and facilitates engineering the features using identifier columns.

We generated various distribution plots to perform a univariate analysis of the numerical data. For this purpose, statistical summaries with the mean, median, and skew were created. First, a data frame with just numerical columns was produced, and then, the mean, median, and skew were calculated from the data frame. Figure 2 depicts the histogram and illustrates the summary of the different attributes present in the training dataset. The figure indicates that the dataset has many skewed variables. For linear models, a uniform distribution is optimal. The target variable in the dataset is not normally distributed.

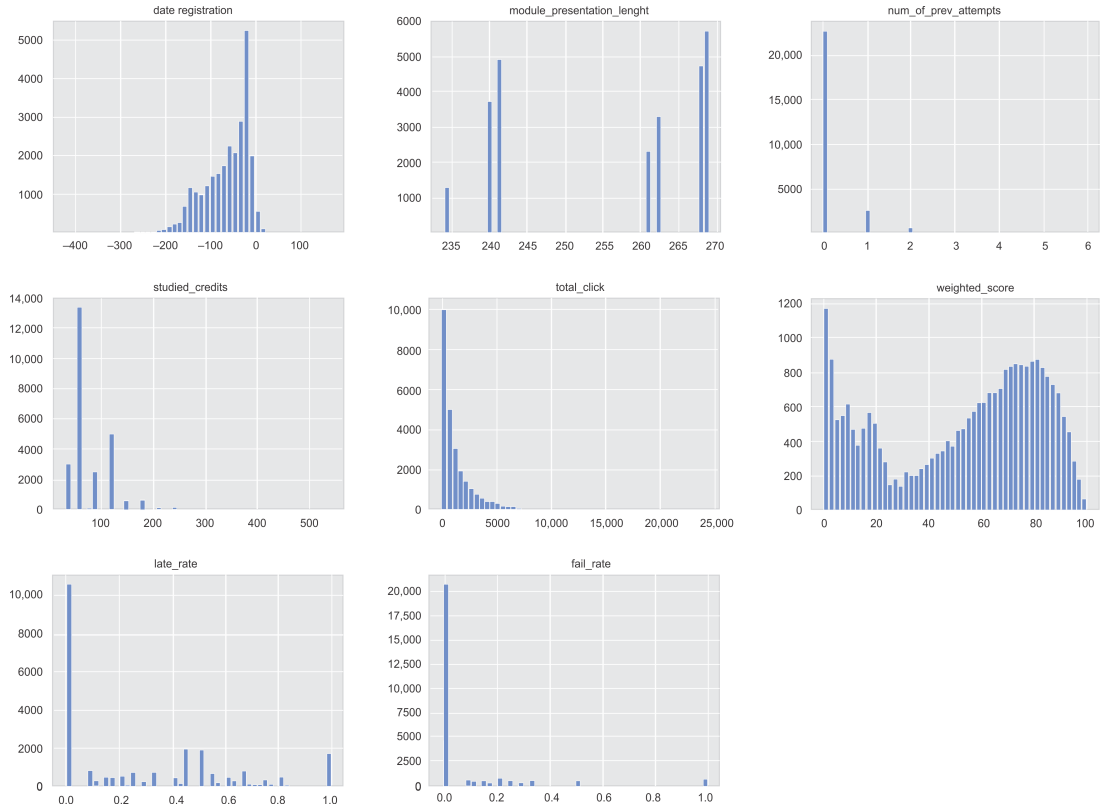

Figure 2. Summary of the attributes in the dataset.

To see the correlation between different features and the target variable, the correlation matrix was generated. Figure 3 shows the correlation heat map of different features. The target variable is the weighted score in this case, which does not have any outliers, but is not normally distributed. The correlation matrix shows that there is some correlation in the features or among the variables in the datasets. It is also relevant to see the correlation between features and the target variable, i.e., the weighted score. Figure 4 shows the linear correlations between features and the target variable. The correlation matrix with the target variable reveals that some variables are very closely related to the target variable, while there are also a few variables that depict a weak correlation with the target variable. There are also some negatively correlated features present in the dataset. It was observed that the weighted score has a high correlation with the total number of clicks. This simply means that, if the students are highly engaged with the portal, then the result percentage should be better and vice versa. It can also be seen from the figure that the number of previous attempts has a negative correlation with the weighted score.

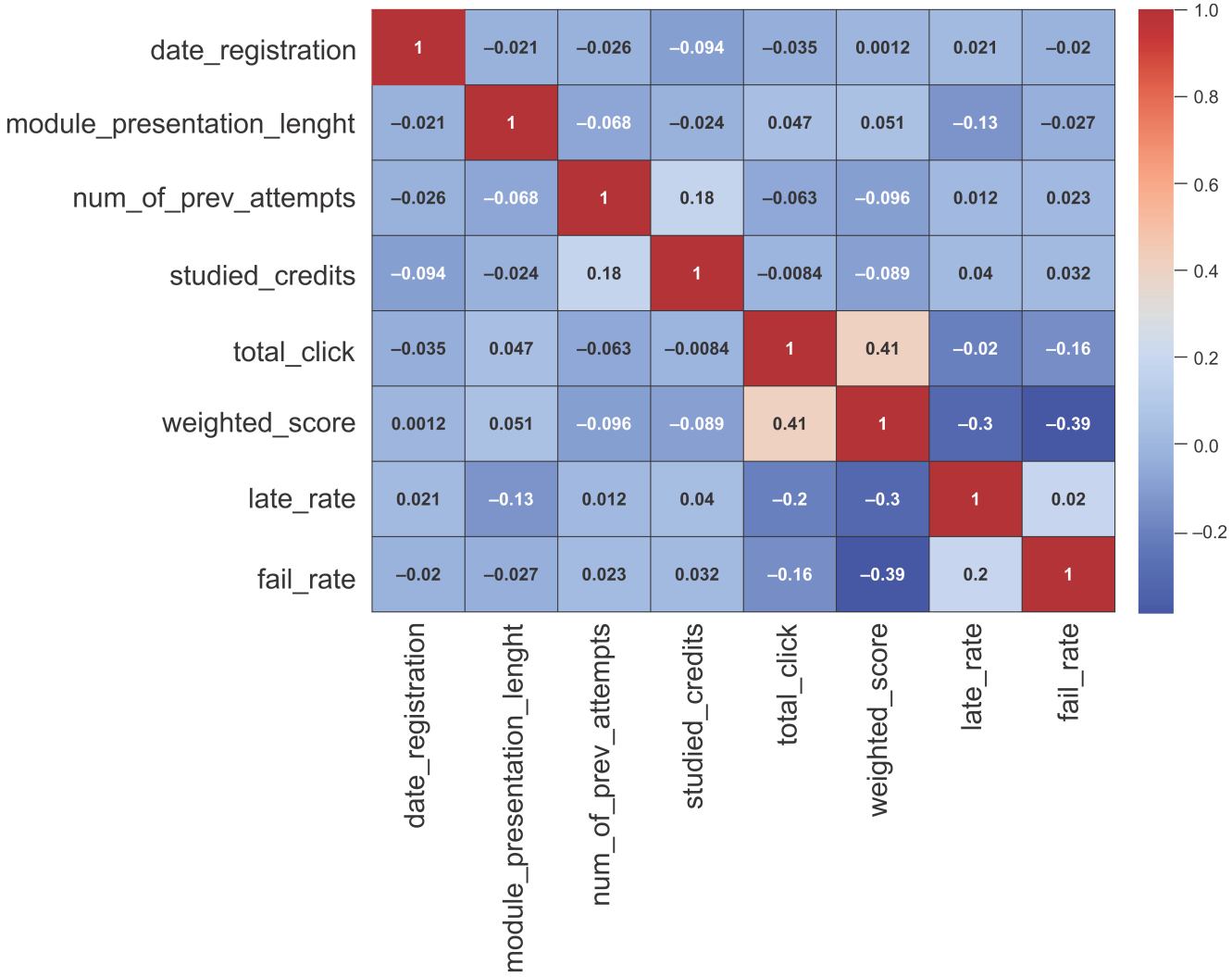

Figure 3. Correlation matrix.

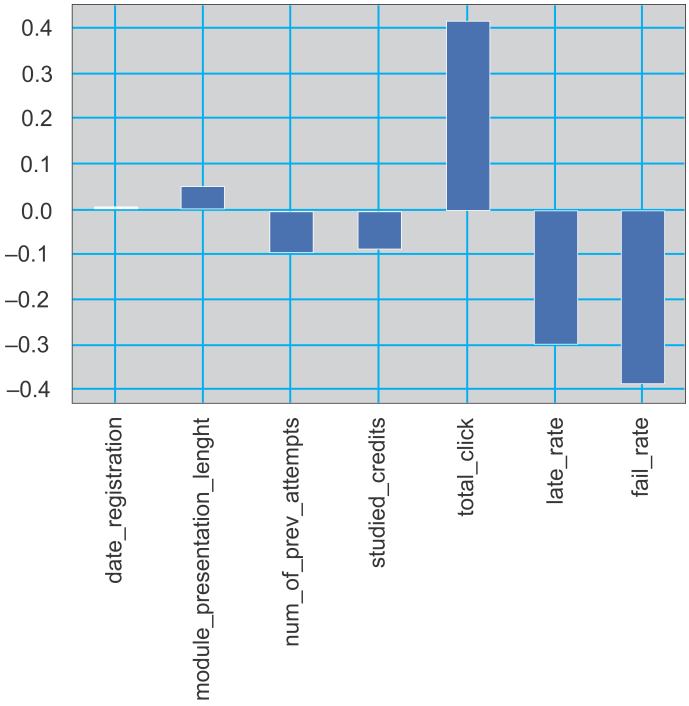

Figure 4. Correlation matrix with target variable.

For the univariate analysis of categorical data, a data frame was created with only categorical data, and various distribution plots were generated. The frequency for each of the attributes was calculated with the help of counting the text property for each label. Figure 5 shows the count plots of various features present in the dataset. These features include the code module, code presentation, gender, and region. It can be seen from the figure that there are 54.9% males present in the dataset, and 45.1% are female. Similarly, the count for other features such as region is shown in the figure.

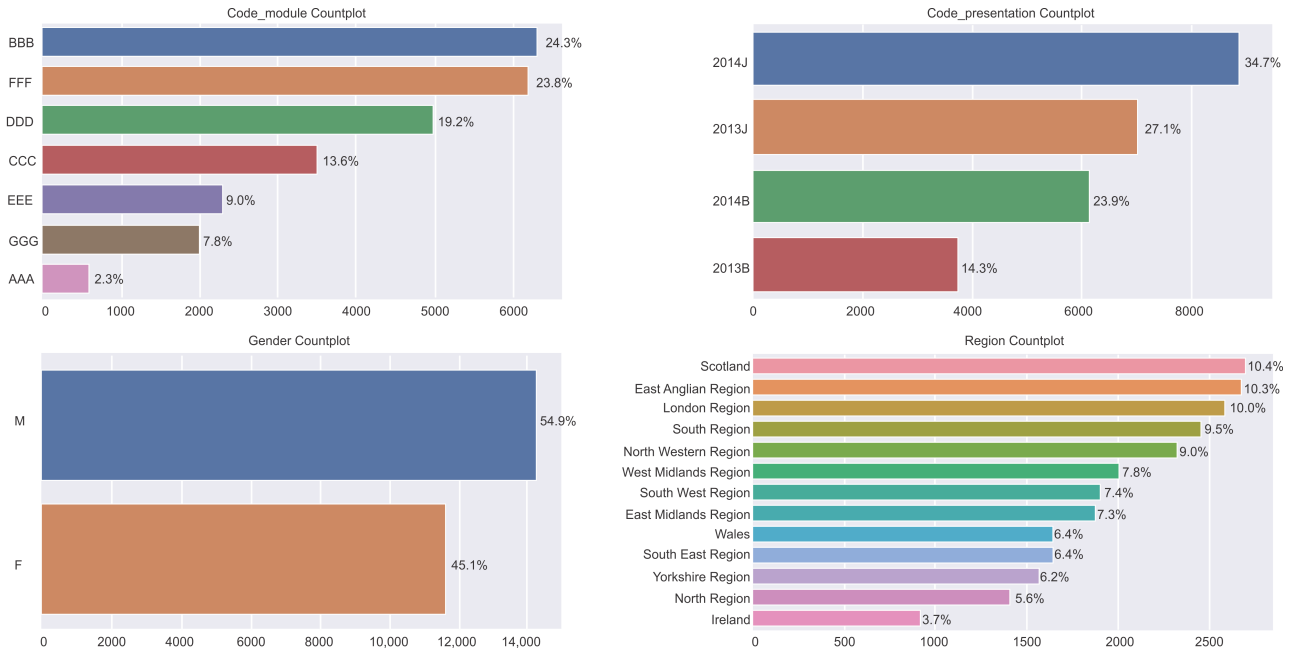

Figure 5. Count plot for categorical variable.

The count plots for other features including the highest education, disability, and age band were also computed, and it was observed that these categories had much fewer

data available. It was possible to merge the two categories of students who had no formal education with those who had post-graduate qualifications with "Lower Than A Level" and "HE Qualification", respectively. Variables with less data did not play an important role in model performance. The same strategy can also be applied to age attributes. As can be seen in Figure 6, the age band of 0–35 had a higher percentage compared to the 35–55 band and those older than 55. Therefore, the latter two bands were merged together.

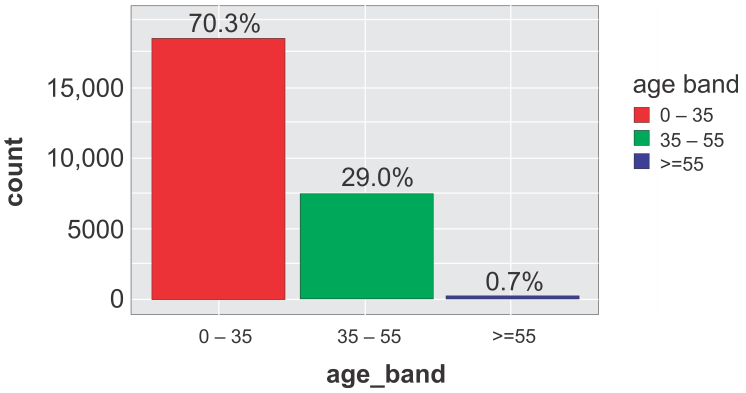

Figure 6. Age attribute.

5.2. Tools and Technologies

The tools and technologies that were used for the implementation and development of the proposed framework mainly consisted of the Python development language. A virtual environment was created for python code development in Anaconda. For machine learning model creation, the Keras framework based on TensorFlow library and other related frameworks and libraries were used. We also used the Pandas library for data manipulation and Matplotlib for the visualization of the data and results.

5.3. Performance Metrics

To evaluate the performance of the proposed framework, several evaluation parameters were used. These evaluation parameters facilitate the understanding of the performance of the overall system, as well as of the proposed model.

Confusion matrix:

A confusion matrix is often used to describe the performance of a classification model on a set of test data that are unseen by the classifier. The confusion matrix is generated in the form of a table, which is relatively easy to understand.

Accuracy:

Accuracy refers to the closeness of a measured value to a standard or known value. This means that the measurements for a given object are close to the known value, but the measurements are far from each other; this then means that the accuracy has no precision.

Precision:

Precision is the quality of being exact. It refers to how close two or more measurements are to each other, regardless of whether those measurements are accurate or not.

Mean absolute error:

The mean absolute error is a statistical measure that is used to measure the errors between observations. These observations are usually paired observations depicting the same phenomenon.

Mean-squared error:

The mean-squared error is also a statistical measure that is used to measure the average of the squares of the errors. It is also known as the mean-squared deviation, which indicates the deviation of the regression line from a set of points.

6. Results and Discussion

This section presents the results of the analysis performed on the datasets discussed in the previous section. As mentioned earlier, three types of analysis tasks were performed. One was regression analysis, and the others were classification and predictive analyses. For each type, multiple algorithms were applied. They are detailed and their results are presented in the subsequent sections.

6.1. Regression Analysis Results

Several regression analysis techniques were employed to view and analyze the underlying dataset and its properties. To prepare the dataset for regression analysis, the training set and the testing set were first separated. All columns which did not play a role in the regression analysis were removed. The resultant dataset was encoded and scaled as mentioned in the previous sections. A column transformer was also applied to all the features. To make the model more efficient and robust, especially against outliers, we used a robust scaler through the inbuilt libraries, which helped scale the data to an interquartile range.

Firstly, the linear relationships between the target variable and the features were predicted. The linear regression is a good choice to see if there is any relationship between features and the target variable or not. Furthermore, the results of linear regression can also indicate the distribution of the features. A bad outcome of linear regression is illustrative of the fact that the features are not linearly distributed.

The following four performance parameters were calculated to measure the performance of linear regression: MSE, RMSE, R2, and adjusted R2. The RMSE turned out to be 23.905, and the adjusted R2 was 0.35, indicating that the model performed poorly on the dataset. This was because the values were not uniformly distributed and there was no linear relationship between the features and the target variable.

When tested with cross-validation, the performance of the model ended with a mean of 23.94 and a standard deviation of 0.24. As mentioned above, stratified k-fold cross-validation was exploited in this research. The mean and standard deviation showed that linear regression was not a good approach to use for this type of dataset.

In order to validate the result inferred from the linear regression, we also tested our hypothesis with lasso regression. Lasso regression is very similar to and a modification of the linear regression, in which some features are dropped depending on the coefficients of those features. If the feature coefficients are low, then they can be dropped.

The results of the lasso regression also validated our previous results, with an RMSE of 23.90 and an R2 with 0.35. When tested with cross-validation, the performance of the model showed a mean of 23.94 and a standard deviation of 0.24 and did not show any improvement. This satisfied our initial assumption that the features were not linearly distributed and there was little multicollinearity present in the features.

It is clear that linear models are unable to generate any analytics from the underlying datasets due to the nonuniform distribution of features and to the limited multicollinearity. Therefore, it is better to test using nonlinear regression models such as the support vector regressor. A support vector regressor (SVR) is a nonlinear regression model that tends to minimize the coefficients instead of minimizing the squared error.

The support vector regressor model ended up with an RMSE of 23.109 and an adjusted R2 score of 0.395. After cross-validation, the performance of the model indicated a mean of 23.417 and a standard deviation of 0.0935 and showed no improvement. The results indicate some improvement while using nonlinear models, as compared to linear models. However, the improvement was still not significant for the SVR model.

To further improve the results, a more complex model known as the gradient boosting regressor was tested on the dataset. The gradient boosting regressor model is a predictive model and is formed by a combination of many weak predictive models that are put together. Due to its unity, it is expected to provide much better results.

The gradient boosting regressor model ended up with an RMSE of 17.714 and an adjusted R2 score of 0.6446. When tested with cross-validation, the performance of the model ended up with a mean of 18.398 and a standard deviation of 0.1973 and showed some improvement: an RMSE of 18.4 for cross-validation and an adjusted R2 score being one point lower as compared to previous results.

To further analyze the data and to record an improvement in the results, the data were considered as a prediction modeling problem, where we were not looking to group the students based on their performance, but to predict their performance based on their commitment to a specific course, their performance during the whole course, and their final grades.

The assessment of the students recorded in the assessment table describes the knowledge of the students for each course. It also contains the grades of the students, which can be very helpful for a prediction problem. We also used the pass rate based on the assessment of the student, as well as their status and participation. The data inferred from the virtual learning environment were used as an indicator to judge the contact hours of the students with the subjects and the quality of their study. Since the pass count was higher compared to the other labels, the least represented cases were examined more closely.

The prediction modeling problem was designed using three different types of models, including logistic regression, linear discriminant analysis (LDA), and random forest. The confusion matrix for the logistic regression showed that the proposed model was 87% precise for both the distinction and fail classes (Figure 7). However, the model was more precise for pass, with a percentage of 89%. The recall and f1-scores are also indicated accordingly.

	precision	recall	f1-score	support
Distinction	0.87	0.79	0.83	273
Fail	0.87	0.70	0.77	204
Pass	0.89	0.95	0.92	1008
accuracy			0.88	1485
macro avg	0.87	0.81	0.84	1485
weighted avg	0.88	0.88	0.88	1485

Figure 7. Confusion matrix for logistic regression.

In the case of linear discriminant analysis, the confusion matrix for the proposed model was 80% for the distinction class and 77% for fail (Figure 8). However, the model was more precise for pass, with a percentage of 92%. The recall and f1-scores are also indicated accordingly.

	precision	recall	f1-score	support
Distinction	0.80	0.86	0.83	273
Fail	0.77	0.79	0.78	204
Pass	0.92	0.90	0.91	1008
accuracy			0.87	1485
macro avg	0.83	0.85	0.84	1485
weighted avg	0.88	0.87	0.87	1485

Figure 8. Confusion matrix for LDA.

For the case of random forest, the confusion matrix showed that the proposed model was 90% and 87% precise for classes of distinction and failure, respectively, as shown in Figure 9. However, the model was more precise for pass, with a percentage of 91%. The recall and f1-scores are also indicated accordingly.

	precision	recall	f1-score	support
Distinction	0.90	0.82	0.85	273
Fail	0.87	0.78	0.82	204
Pass	0.91	0.95	0.93	1008
accuracy			0.90	1485
macro avg	0.89	0.85	0.87	1485
weighted avg	0.90	0.90	0.90	1485

Figure 9. Confusion matrix for random forest.

6.2. Classification Results

Three different types of classification techniques were used to see and analyze the underlying dataset and its properties. This analysis helped to generate the required insights from the dataset. To prepare the dataset for the classification task, the training set and testing set were first separated. The associated data that did not play any role in the classification task were removed. The resultant dataset was then encoded and scaled, as was done for the regression analysis. The column transformer was also applied to all the features.

The decision tree classifier was first used to perform the simplest kind of classification. The minimum sample leaf size was selected to be 15 with a minimum sample split size of 10. The maximum feature value was selected as eight. We calculated the following three performance parameters to measure the performance of the decision tree classifier: mean, standard deviation, and performance scores. The mean turned out to be 21.22, and the standard deviation was 0.23, which indicates that the model was not performing well on the dataset.

To further analyze the dataset, a random forest classifier was chosen to generate insights from the dataset. The parameters of the random forest classifier were selected to be almost the same as those of the decision tree classifier. The minimum leaf sample size was selected as 15 with a minimum sample split size of 10. The maximum features were again selected to be eight, with the total number of estimators equal to 20.

However, the performance of both the decision tree classifier and random forest classifier was not up to the mark. To further improve the results, support vector machines were then utilized, but encoding the values for the support vector machines required scaling. To make the SVM more efficient and robust, especially against outliers, we again used a robust scaler through the inbuilt libraries, which helped to scale the data to an interquartile range. The gamma value for the SVC classifier was set to auto.

The comparison of the above-mentioned classifiers tested on the dataset revealed that the support vector machine was the best-performing classifier, with an accuracy of 78% and a standard deviation of 0.002. Furthermore, in comparison to the others, the support vector classifier model indicated less variance between the scores during cross-validation, as shown by the lower standard deviation.

We also treated the underlying dataset as a prediction modeling problem and predicted the performance of the students based on the commitment of the students to a specific course, their performance for the whole course, and their final grades and used all the features that we engineered together and explained in the feature engineering section. To perform the analysis, we used an artificial neural network with the most optimized parameters.

A sequential model with three dense layers and a dropout layer was developed. The first two dense layers and the dropout layer used the rectified linear unit (ReLU) as the activation function. The last dense layer operated with the sigmoid as the activation function to perform the classification. The loss function was selected to be binary cross-entropy, and Adam was the optimization function. The network was trained for almost 200 epochs on the training dataset to achieve an acceptable value of loss. Figure 10 shows the training and validation loss over multiple epochs during training.

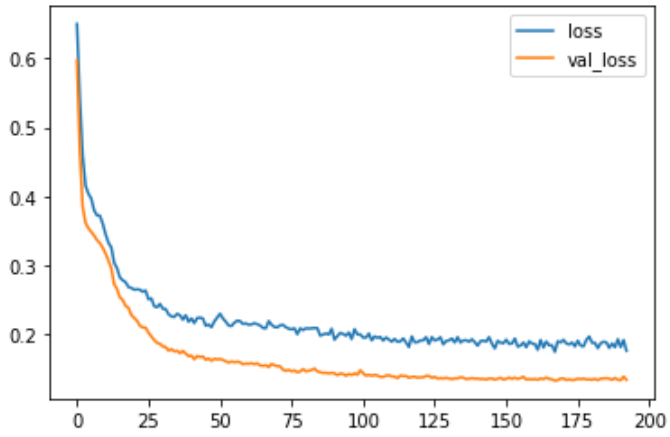

Figure 10. Performance of the neural network.

The confusion matrix for the artificial neural network as shown in Figure 11 indicates that the proposed model was 95% precise for both classes. The recall and f1-scores are also indicated accordingly.

The proposed machine-learning-based interpretable framework can be helpful for staff and teachers in a number of ways. For example, the proposed framework can be used to estimate the final results of a particular student based on his/her intermediary marks. A teacher could know in advance from the output of the framework whether a student is going to pass or fail in the module, and thus, he/she can take appropriate measures well in time to improve the students' performance. The proposed framework can also be used for the rapid classification of a large number of successful and at-risk students automatically. This will reduce the manual burden and processing time of both staff and teachers. Another major advantage of the proposed machine-learning-based interpretable framework is the identification of factors that are playing a major role in the success or failure of a student. The staff and teachers can take appropriate actions based on the identified factors for further academic improvement of students.

	precision	recall	f1-score	support
0	0.94	0.66	0.77	202
1	0.95	0.99	0.97	1283
accuracy			0.95	1485
macro avg	0.94	0.83	0.87	1485
weighted avg	0.95	0.95	0.94	1485

Figure 11. Confusion matrix for the neural network.

7. Conclusions and Future Work

In this research study, we proposed a machine-learning-based interpretable framework for e-learning platforms with the main purpose of helping both students and teachers improve their academic performance. E-learning platforms generate a large amount of data related to student behavior that can be efficiently processed and analyzed using intelligent techniques to produce useful insights that can help improve student performance. The proposed system incorporates predictive models, as well as regression and classification models to analyze multiple aspects of student performance. The classification models systematized normal and at-risk students based on their academic performance, with high precision and accuracy. The regression analysis determined the inherent linear and nonlinear relationships between the students' academic outcomes and the performance

indicative features. The proposed framework helped determine and analyze the underlying dataset and its properties. This analysis helped to generate useful insights about the performance of students, such as identifying the factors impacting the progress of successful students or the performance of the students who are struggling in their courses and are at risk of failing.

In the future, we would like to perform more detailed analyses on a much larger dataset with deep-learning-based approaches. The proposed framework will be further enhanced to incorporate more data, and the underlying GPU-based infrastructure will be scaled to facilitate distributed machine learning training and deployment.

Author Contributions: Conceptualization, I.G.; Data curation, H.G.; Formal analysis, I.G. and R.O.; Methodology, M.U.Y. and H.G.; Resources, I.G.; Software, M.U.Y. and R.O.; Supervision, M.C.; Validation, M.U.Y. and M.C.; Visualization, M.U.Y.; Writing—original draft, I.G., M.C. and R.O.; Writing—review & editing, I.G., M.U.Y., M.C. and H.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sabi, H.M.; Uzoka, F.M.E.; Langmia, K.; Njeh, F.N. Conceptualizing a model for adoption of cloud computing in education. *Int. J. Inf. Manag.* **2016**, *36*, 183–191. [CrossRef]
2. Ramírez-Montoya, M.S.; Andrade-Vargas, L.; Rivera-Rogel, D.; Portuguez-Castro, M. Trends for the Future of Education Programs for Professional Development. *Sustainability* **2021**, *13*, 7244. [CrossRef]
3. Herodotou, C.; Rienties, B.; Hlosta, M.; Boroowa, A.; Mangafa, C.; Zdrahal, Z. The scalable implementation of predictive learning analytics at a distance learning university: Insights from a longitudinal case study. *Internet High. Educ.* **2020**, *45*, 100725. [CrossRef]
4. Macfadyen, L.P.; Dawson, S. Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Comput. Educ.* **2010**, *54*, 588–599. [CrossRef]
5. Collberg, C.; Debray, S.; Kobourov, S.; Westbrook, S. Increasing Undergraduate Involvement in Computer Science Research. In Proceedings of the 8th World Conference on Computers in Education (WCCE), Cape Town, South Africa, 4–7 July 2005.
6. Noblitt, L.; Vance, D.E.; Smith, M.L.D. A comparison of case study and traditional teaching methods for improvement of oral communication and critical-thinking skills. *J. Coll. Sci. Teach.* **2010**, *39*, 26–32.
7. Li, Y.; Nishimura, N.; Yagami, H.; Park, H.S. An Empirical Study on Online Learners' Continuance Intentions in China. *Sustainability* **2021**, *13*, 889. [CrossRef]
8. Portillo, J.; Garay, U.; Tejada, E.; Bilbao, N. Self-Perception of the Digital Competence of Educators during the COVID-19 Pandemic: A Cross-Analysis of Different Educational Stages. *Sustainability* **2020**, *12*, 128. [CrossRef]
9. Bowles, M. Learning to E-Learn Project: Rediscovering the benefits of e-learning. *Malays. Online J. Instr. Technol.* **2005**, *2*, EJ.
10. Mothibi, G. A Meta-Analysis of the Relationship between E-Learning and Students' Academic Achievement in Higher Education. *J. Educ. Pract.* **2015**, *6*, 6–9.
11. Abulibdeh, E.S.; Hassan, S.S.S. E-learning interactions, information technology self efficacy and student achievement at the University of Sharjah, UAE. *Australas. J. Educ. Technol.* **2011**, *27*, 1014–1025. [CrossRef]
12. Riahi, G. E-learning systems based on cloud computing: A review. *Procedia Comput. Sci.* **2015**, *62*, 352–359. [CrossRef]
13. Baepler, P.; Murdoch, C.J. Academic analytics and data mining in higher education. *Int. J. Scholarsh. Teach. Learn.* **2010**, *4*, 17. [CrossRef]
14. Bin Mat, U.; Buniyamin, N.; Arsad, P.M.; Kassim, R. An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention. In Proceedings of the 2013 IEEE 5th Conference on Engineering Education (ICEED), Kuala Lumpur, Malaysia, 4–5 December 2013; pp. 126–130.
15. Huebner, R.A. A Survey of Educational Data-Mining Research. *Res. High. Educ. J.* **2013**, *19*, 1–13.
16. Rodrigues, M.W.; Isotani, S.; Zárate, L.E. Educational Data Mining: A review of evaluation process in the e-learning. *Telemat. Inform.* **2018**, *35*, 1701–1717. [CrossRef]

17. Romero, C.; Ventura, S. Educational data mining and learning analytics: An updated survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2020**, *10*, e1355. [CrossRef]
18. Cassells, L. The effectiveness of early identification of 'at risk' students in higher education institutions. *Assess. Eval. High. Educ.* **2017**, *43*, 515–526. [CrossRef]
19. Tran, D.P.; Nguyen, G.N.; Hoang, V.D. Hyperparameter Optimization for Improving Recognition Efficiency of an Adaptive Learning System. *IEEE Access* **2020**, *8*, 160569–160580. [CrossRef]
20. Gardner, J.; Brooks, C. Student success prediction in MOOCs. *User Model. User-Adapt. Interact.* **2018**, *28*, 127–203. [CrossRef]
21. Liao, S.N.; Zingaro, D.; Thai, K.; Alvarado, C.; Griswold, W.G.; Porter, L. A robust machine learning technique to predict low-performing students. *ACM Trans. Comput. Educ.* **2019**, *19*, 18. [CrossRef]
22. Gray, C.C.; Perkins, D. Utilizing early engagement and machine learning to predict student outcomes. *Comput. Educ.* **2019**, *131*, 22–32. [CrossRef]
23. Asif, R.; Merceron, A.; Ali, S.A.; Haider, N.G. Analyzing undergraduate students' performance using educational data mining. *Comput. Educ.* **2017**, *113*, 177–194. [CrossRef]
24. Polyzou, A.; Karypis, G. Feature Extraction for Classifying Students Based on Their Academic Performance. In Proceedings of the 11th International Educational Data Mining Society, Buffalo, NY, USA, 15–18 July 2018.
25. Ray, S.; Saeed, M. Applications of Educational Data Mining and Learning Analytics Tools in Handling Big Data in Higher Education. In *Applications of Big Data Analytics: Trends, Issues, and Challenges*; Springer: Cham, Switzerland, 2018; pp. 135–160. [CrossRef]
26. Brohi, S.N.; Pillai, T.R.; Kaur, S.; Kaur, H.; Sukumaran, S.; Asirvatham, D. Accuracy Comparison of Machine Learning Algorithms for Predictive Analytics in Higher Education. In *International Conference for Emerging Technologies in Computing*; Springer: Cham, Switzerland, 2019; Volume 285, pp. 254–261. [CrossRef]
27. Saqr, M.; Fors, U.; Tedre, M. How learning analytics can early predict under-achieving students in a blended medical education course. *Med. Teach.* **2017**, *39*, 757–767. [CrossRef] [PubMed]
28. Chaichumpa, S.; Temdee, P. Multi-agents platform for mobile learning using objective distance based personalisation method. *Int. J. Mob. Learn. Organ.* **2018**, *12*, 293–310. [CrossRef]
29. Baashar, Y.; Alkawsi, G.; Ali, N.; Alhussian, H.; Bahbouh, H.T. Predicting student's performance using machine learning methods: A systematic literature review. In Proceedings of the International Conference on Computer and Information Sciences: Sustaining Tomorrow with Digital Innovation, ICCOINS, Kuching, Malaysia, 13–15 July 2021; pp. 357–362. [CrossRef]
30. Chen, F.; Cui, Y. Utilizing Student Time Series Behaviour in Learning Management Systems for Early Prediction of Course Performance. *J. Learn. Anal.* **2020**, *7*, 1–17. [CrossRef]
31. Saleem, F.; Ullah, Z.; Fakieh, B.; Kateb, F. Intelligent Decision Support System for Predicting Student's E-Learning Performance Using Ensemble Machine Learning. *Mathematics* **2021**, *9*, 2078. [CrossRef]
32. Al Duhayyim, M.; Newbury, P. Concept-based and Fuzzy Adaptive E-learning. In Proceedings of the 2018 3rd International Conference on Information and Education Innovations, London, UK, 30 June–2 July 2018; pp. 49–56. [CrossRef]
33. Matazi, I.; Bennane, A.; Messoussi, R.; Touahni, R.; Oumaira, I.; Korchiyne, R. Multi-Agent System Based on Fuzzy Logic for E-Learning Collaborative System. In Proceedings of the International Symposium on Advanced Electrical and Communication Technologies, ISAECT 2018–Proceedings, Rabat, Morocco, 21–23 November 2018. [CrossRef]
34. Noama, K.M.G.; Khalid, A.; Muharram, A.A.; Ahmed, I.A. Improvement of E-learning Based via Learning Management Systems (LMS) Using Artificial Neural Networks. *Asian J. Res. Comput. Sci.* **2019**, *4*, 1–9. [CrossRef]
35. McGill, T.J.; Klobas, J.E. A task–technology fit view of learning management system impact. *Comput. Educ.* **2009**, *52*, 496–508. [CrossRef]
36. Khanal, S.S.; Prasad, P.; Alsadoon, A.; Maag, A. A systematic review: Machine learning based recommendation systems for e-learning. *Educ. Inf. Technol.* **2020**, *25*, 2635–2664. [CrossRef]
37. Chatti, M.A.; Dyckhoff, A.L.; Schroeder, U.; Thüs, H. A reference model for learning analytics. *Int. J. Technol. Enhanc. Learn.* **2012**, *4*, 318–331. [CrossRef]
38. Masud, M.A.H.; Huang, X. A novel approach for adopting cloud-based e-learning system. In Proceedings of the 2012 IEEE/ACIS 11th International Conference on Computer and Information Science, Shanghai, China, 30 May–1 June 2012; pp. 37–42.
39. Azzi, I.; Jeghal, A.; Radouane, A.; Yahyaouy, A.; Tairi, H. A robust classification to predict learning styles in adaptive E-learning systems. *Educ. Inf. Technol.* **2020**, *25*, 437–448. [CrossRef]
40. Eggen, T.; Straetmans, G. Computerized adaptive testing for classifying examinees into three categories. *Educ. Psychol. Meas.* **2000**, *60*, 713–734. [CrossRef]
41. Lykourentzou, I.; Giannoukos, I.; Nikolopoulos, V.; Mpardis, G.; Loumos, V. Dropout prediction in e-learning courses through the combination of machine learning techniques. *Comput. Educ.* **2009**, *53*, 950–965. [CrossRef]
42. Khamparia, A.; Pandey, B. SVM and PCA based learning feature classification approaches for e-learning system. *Int. J. Web-Based Learn. Teach. Technol. (IJWLTT)* **2018**, *13*, 32–45. [CrossRef]
43. Khamparia, A.; Singh, S.K.; Luhach, A.K.; Gao, X.Z. Classification and analysis of users review using different classification techniques in intelligent e-learning system. *Int. J. Intell. Inf. Database Syst.* **2020**, *13*, 139–149. [CrossRef]

44. Sheeba, T.; Krishnan, R. Prediction of student learning style using modified decision tree algorithm in e-learning system. In Proceedings of the 2018 International Conference on Data Science and Information Technology, Singapore, 20–22 July 2018; pp. 85–90.
45. Şuşnea, E. Using artificial neural networks in e-learning systems. *UPB Sci. Bull. Ser. C* **2010**, *72*, 91–100.
46. Chang, H.S.; Hsu, H.J.; Chen, K.T. Modeling Exercise Relationships in E-Learning: A Unified Approach. In Proceedings of the International Conference on Educational Data Mining (EDM), Madrid, Spain, 26–29 June 2015; pp. 532–535.
47. Yang, S.J.; Lu, O.H.; Huang, A.Y.; Huang, J.C.; Ogata, H.; Lin, A.J. Predicting students' academic performance using multiple linear regression and principal component analysis. *J. Inf. Process.* **2018**, *26*, 170–176. [CrossRef]
48. Rajalaxmi, R.; Natesan, P.; Krishnamoorthy, N.; Ponni, S. Regression model for predicting engineering students academic performance. *Int. J. Recent Technol. Eng.* **2019**, *7*, 71–75.
49. Thompson, E.D.; Bowling, B.V.; Markle, R.E. Predicting student success in a major's introductory biology course via logistic regression analysis of scientific reasoning ability and mathematics scores. *Res. Sci. Educ.* **2018**, *48*, 151–163. [CrossRef]
50. Moubayed, A.; Injadat, M.; Shami, A.; Lutfiyya, H. Relationship between student engagement and performance in e-learning environment using association rules. In Proceedings of the 2018 IEEE World Engineering Education Conference (EDUNINE), Buenos Aires, Argentina, 11–14 March 2018; pp. 1–6.
51. Angeline, D.M.D. Association rule generation for student performance analysis using apriori algorithm. *SIJ Trans. Comput. Sci. Eng. Appl. (CSEA)* **2013**, *1*, 12–16. [CrossRef]
52. Govindasamy, K.; Velmurugan, T. Analysis of student academic performance using clustering techniques. *Int. J. Pure Appl. Math.* **2018**, *119*, 309–323.
53. Shovon, M.; Islam, H.; Haque, M. An Approach of Improving Students Academic Performance by using k means clustering algorithm and Decision tree. *arXiv* **2012**, arXiv:1211.6340.
54. Waheed, H.; Hassan, S.U.; Aljohani, N.R.; Hardman, J.; Alelyani, S.; Nawaz, R. Predicting academic performance of students from VLE big data using deep learning models. *Comput. Hum. Behav.* **2020**, *104*, 106189. [CrossRef]
55. Giannakas, F.; Troussas, C.; Voyiatzis, I.; Sgouropoulou, C. A deep learning classification framework for early prediction of team-based academic performance. *Appl. Soft Comput.* **2021**, *106*, 107355. [CrossRef]
56. 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]
57. Kuzilek, J.; Hlosta, M.; Zdrahal, Z. Open university learning analytics dataset. *Sci. Data* **2017**, *4*, 1–8. [CrossRef]